Music to the brain

-

Music to the Brain: investigating auditory scene analysis with polyphonic music

Niels R. Disbergen

The work in this thesis was supported by The Netherlands Organisation for Scientific Research (NWO) Research Talent grant (406-12-126) to Niels R. Disbergen and Elia Formisano, NWO Vici grant (453-12-002) to Elia Formisano, European Union Erasmus Mundus Exchange Scholarship to Niels R. Disbergen, Maastricht University, and operating funds from the Canadian Institutes for Health Research to Robert Zatorre.

Cover design: Niels R. Disbergen
Production: Ipskamp Printing, Enschede
ISBN: 978-94-028-1978-6

Music to the Brain: investigating auditory scene analysis with polyphonic music

DISSERTATION

to obtain the degree of Doctor at Maastricht University, on the authority of the Rector Magnificus Prof. Dr. Rianne M. Letschert, in accordance with the decision of the Board of Deans, to be defended in public on Friday 29 May 2020 at 16:00 hours

by

Niels Robert Disbergen

Supervisors
Prof. Dr. Elia Formisano
Prof. Dr. Robert J. Zatorre, *McGill University, Montreal, Canada*

Co-supervisor
Dr. Giancarlo Valente

Assessment Committee
Prof. Dr. Bernadette Jansma (chair)
Prof. Dr. Elvira Brattico, *Aarhus University Hospital, Aarhus, Denmark*
Dr. Fabrizio Esposito, *Scuola Medica Salernitana, Salerno, Italy*
Dr. Lars Riecke

Contents

”Experience is merely the name men gave to their mistakes”
- Oscar Wilde

1

General Introduction

In the subway station, during rush hour. You are trying to decipher the service announcement while simultaneously suppressing the morning commute playlist playing in your earphones as well as the many buzzing sounds around you. After the announcement ends, you return to your music and immerse yourself again in the ongoing instrument solo. Such a situation exemplifies the typical auditory scene analysis challenges which our auditory system faces on a daily basis. In many cases, the auditory system is presented with multiple sounds entering our ears simultaneously, which are mixed into a single blur of sound waves. These sound waves are transmitted through the ear canal and make the tympanic membrane vibrate. The tympanic vibrations are for the first time analyzed within the cochlea, separating them purely based on their frequency content. In order to perceive a sound of interest (e.g., the service announcement), the auditory system somehow needs to group all the separate frequencies which belonging to the sound of interest again, segregating them from the concurrent frequencies of all the other sounds which are present within the mixture.

This thesis investigates how the auditory system achieves such a seemingly daunting task. More specifically, it examines the neural mechanisms responsible for separating these multiple simultaneous sounds present in the auditory scene. Aside from sound separation, it is possible to combine these same sounds into single merged coherent percepts, which is especially topical in the context of multi-instrument music perception (Fig. 1.1).

The current chapter will provide the theoretical and methodological background for the empirical studies presented in the remainder of this thesis. Auditory scene analysis and music perception concepts are introduced, followed by a discussion of the experimental approaches and analysis methodology employed, and concluding with a general outline of the thesis.

Auditory Scene Analysis

Mechanisms involved in auditory scene analysis (ASA) are essential to hearing in daily circumstances. An elegant framework for this perceptual organization of sounds has been comprehensively described in Albert Bergman's seminal book (Bregman, 1990). In general, ASA, sometimes referred to as perceptual grouping (Darwin, 1981), describes those mechanisms underlying the organization of acoustic sources comprised in sound mixtures into distinct auditory events (for reviews, see Bregman, 1990; Ciocca, 2008). A sequential organization of auditory events can be referred to as an auditory stream, or in short stream; note that in this work both terms will be employed interchangeably. A stream is, therefore, the perceptual representation from a series of sounds which is perceived as one single (continuous) entity, invariantly so with respect to changes in acoustics and background noise (e.g., Griffiths & Warren, 2004). Within typically

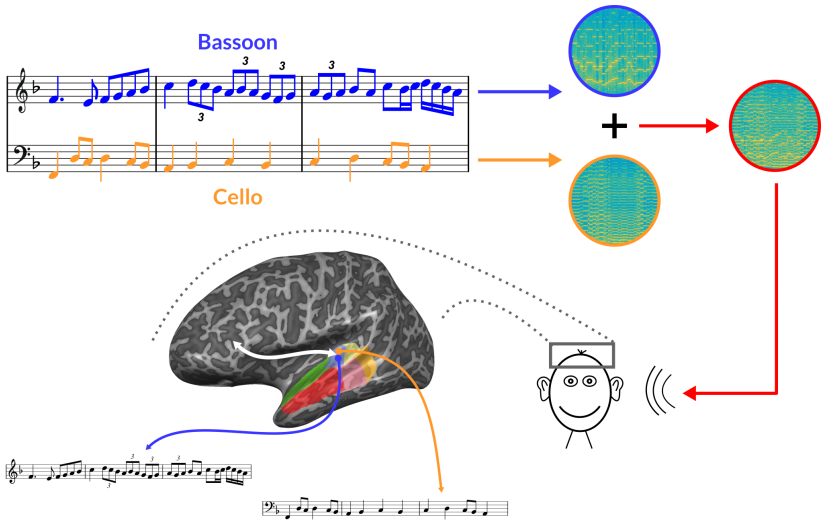


Figure 1.1: Schematic Representation of Music Scene Analysis. Listeners are presented with two-instrument polyphonic music played by a Bassoon and a Cello. Both of these instruments are blended into a single soundwave when entering the ear. The brain performs a multitude of computations upon this signal, among which the interaction between temporal auditory and frontal (attentional) regions is of great importance in order to once again separate the mixed sound signal into the individual melodies played by the distinct instruments.

highly dynamic auditory scenes, such invariant representations provide the necessary mechanisms to capture the acoustic signal variance and allow for the continuous tracking of sounds. The presence of (rudimentary) ASA mechanisms has also been demonstrated for several non-human species, including monkeys, birds, and fishes (Bee & Micheyl, 2008; Fay, 1998, 2000; Hulse et al., 1997; Izumi, 2002).

Due to a stream's definition as a perceptual entity, it is useful to note that streams do not necessarily represent the physical sound sources they originated from. It is in general beneficial for discussion to create a distinction between a sound source and a stream, as in literature these are often employed interchangeably, even though Bregman (e.g., Bregman, 1990) also distinguished between them. A sound source is the physical entity which gives rise to the acoustic pressure waves (e.g., a speaker), while the auditory stream is the percept from a group of consecutive and/or simultaneous sound elements as one coherent item, which does not correspond necessarily with a single sound source.

Perceptual ASA Mechanisms

Stream segregation forms an essential process in ASA, underlying the parcellation of scenes containing multiple spectrally and temporally overlapping sounds into their distinct auditory streams. Auditory streaming mechanisms are classically divided into two distinct types, gen-

eral purpose (*i.e.*, primitive) and schema-based (e.g., Bregman, 1990; Ciocca, 2008; Shamma & Micheyl, 2010). General purpose mechanisms are responsible for the on-line acoustical analysis of spectral and temporal relations between scene elements and are hypothesized to be mostly bottom-up driven and pre-attentive. Sequential grouping based on these general-purpose mechanisms could, for example, be performed by frequency similarity or temporal proximity. Simultaneous matching could employ, among others, spectral regularities, coherent spectral modulations, common onset, or spatial cues (e.g., Bregman, 1990; Ciocca, 2008). Schema-based mechanisms, on the contrary, are higher-level mechanisms which are mostly under top-down influence and considered to be domain specific, for example applying to music. Schemas can strongly vary with regard to their complexity and are modulated by previous exposure and learning. Schemas are hypothesized to perform, for example, both the matching of ongoing sound features with existing categories, as well as being involved in the generation of perceptual attributes such as pitch or loudness by employing dynamic combinations of lower-level features, such as spectral and temporal regularities. A classic example of a higher-level schema-based mechanism is the phonemic restoration effect in speech, where a physically omitted phoneme is restored by the brain and the word is nonetheless perceived as continuous (Warren, 1970). Analysis of naturalistic auditory scenes involves a dynamic interaction between these general-purpose and schema-based mechanisms, where the segregation and possible combination of sounds can be further influenced by the listener's attention (e.g., Elhilali et al., 2009b).

Top-down and bottom-up mechanisms refer to general perceptual system operations, not to be equated to the schema-based and general-purpose mechanisms, respectively, upon which they operate. Bottom-up mechanisms represent the input-driven information flow typically origination from the senses, while top-down refers to the influences higher-level processing stages can exert onto the ongoing bottom-up driven stimulus analysis. Formation of auditory streams can, in general, be influenced by various top-down mechanisms, among which are, most fitting so to the current thesis, attention and extensive training (e.g., Bregman, 1990; Kraus & Chandrasekaran, 2010; Snyder et al., 2012; Sussman, 2017). Attentive mechanisms in particular may influence the analysis of auditory scenes at many different levels, for example biasing towards the employment of specific grouping mechanisms, or the modulation of specific sound feature representations (see Fig. 1.2; (Sussman, 2017)).

The current work focuses on top-down attentive processes and did not implement any manipulation of physical top-down cues which are additionally capable of influencing streaming towards segregation or integration, even though they are an important part of (music) ASA (e.g., McAdams & Bregman, 1979; Bregman, 1990). For example, when performing stream segregation of two interleaved melodies, it has been shown that prior perception of the target sequence results in better subsequent segregation performance, while after performing a frequency-transposition

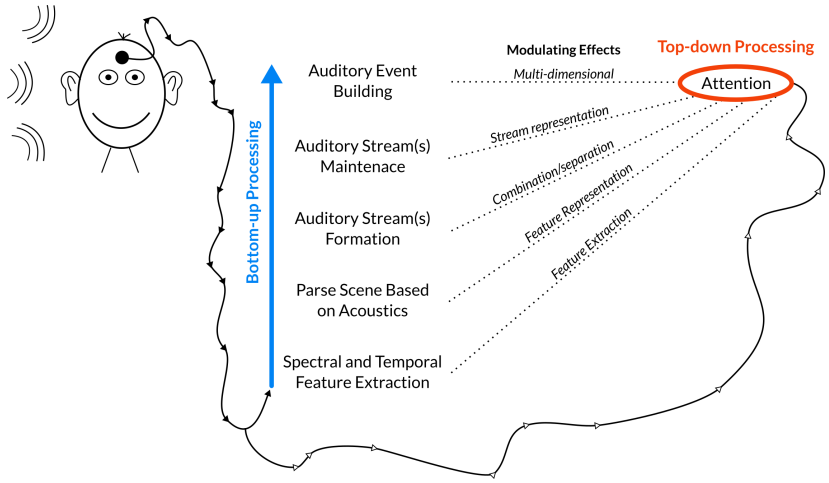


Figure 1.2: Bottom-up and Top-down Processing Mechanisms in ASA. Generalized perceptual processing mechanisms can be divided in a bottom-up stimulus driven system (blue arrow) as well as a top-down attention-driven system (red circle). Top-down attentive processes are capable of modulating the ongoing bottom-up processing during a multitude of different processing levels, ranging from 'low-level' feature extractions up until high-level multi-dimensional auditory event building.

of this same melody these effects were reduced (Bey & McAdams, 2003). For an in-depth discussion of ASA in general, along with the low-level features involved in both the grouping and segregation of sounds, see, for example, Bregman (1990) and Ciocca (2008).

Neural ASA Mechanism

Since the publication of Bregman's (1990) work, a large body of additional psychophysical research on ASA has been performed, even though the underlying neural mechanisms have been studied to a much more limited extend. The majority of neuroscientific investigations have employed sounds presented in isolation or within elementary auditory scenes, such as tones in noise or alternating tone sequences. Classically, ASA mechanisms have been studied by using variations of the ABA paradigm (Bregman & Campbell, 1971; van Noorden, 1971). Within this setup, two pure tones of different frequencies are presented alternating rapidly. These sequences are perceived as one or two streams, dependent, among others, on their acoustic differences. Common tone-sequence manipulations include differences in their frequency or timbre (e.g., van Noorden, 1977; Bregman, 1990). These streaming effects are often not instantaneous and build up over a time-course of several seconds, the neuronal adaptations to which have been observed along the auditory pathway as early as in the cochlear nucleus (e.g., Bregman, 1978; Pressnitzer et al., 2008).

In the context of explaining streaming effects for alternating tone-sequences, population separation models suggest that the individual streams are represented by distinct neuronal popu-

lations (e.g., Fishman et al., 2012; Micheyl et al., 2007). The temporal coherence model adds to this the subsequent importance of synchronicity cues in further sharpening the population separation, which facilitates sound-feature extraction and grouping (Elhilali et al., 2009a; Shamma et al., 2010). An additional line of reasoning adds modules for predictive coding (e.g., Winkler et al., 2012), where higher-level cortices perform continues predictions based on previous experience and incoming sensory signals, stressing the importance of top-down mechanisms alongside the previously established bottom-up mechanisms. Streaming tasks have been shown to also involve large sections of cortex beyond the classical auditory areas, including frontal, temporal, and parietal regions (Dykstra et al., 2011). Sources for the top-down control of auditory cortex involved in stream formation may also include regions in the intraparietal sulcus or the superior temporal sulcus (Belin et al., 2000; Binder et al., 2000; Cusack, 2005; Teki et al., 2011).

Emergence of auditory percepts from incoming acoustical information spans a large array of neural mechanisms, as diverse as early feature extraction until higher order cognitive processes such as schema based matching (e.g. Alain & Bernstein, 2008; Bregman, 1990; Ciocca, 2008). A debate remains as to whether stream segregation requires attention, or it is a fully pre-attentive bottom-up driven mechanism (e.g., Carlyon et al., 2001; Lakatos et al., 2013; Macken et al., 2003; Winkler et al., 2003). The necessity of attention for stream maintenance mechanisms per se is much less debated (e.g., Sussman et al., 2007), even though it remains elusive as to exactly where and when in the brain it is capable of influencing the ongoing sound processing. This thesis will focus mostly on the mechanisms involved in the analysis of complex sounds, *id est* those containing two or more temporally overlapping frequency components, unless otherwise specified.

Analysis of complex sound scenes requires some form of multi-level analysis, incorporating in a task-dependent fashion a dynamic interaction between stimulus-driven bottom-up and attentively-guided top-down mechanism (e.g., Sussman, 2017). Early stimulus analysis windows probably mostly reflect the processing of physical features important for their grouping, for example spectral information, representing a first lower-level stimulus abstraction which forms the basis for the eventual formation of perceptual streams. The more stimulus analysis progresses, the more activation patterns potentially start to represent their perceptual attributes as opposed to direct physical acoustic relationships. One possible approach to implementing such an analysis would be a multi-level mechanism, which entails a first bottom-up driven analysis sweep leading to the initial segregation of stimuli, presuming sufficient physical differences exist between them. Following this initial analysis sweep, attention and other top-down processes could interact with the ongoing stimulus representations, potentially modulating them at these lower levels. Under common processing circumstances, the attentively-driven mechanisms probably operate on the higher-level stream representations as opposed to their lower-level acoustic vari-

ants. Even though the system could gain access to these earlier, potentially subcortical, stimulus representations at rapid time-scales in a reverse hierarchical fashion (Nahum et al., 2008). Modulations of early analysis windows, albeit subtle, have been observed at the level of the inferior colliculus and auditory cortex (Poghosyan & Ioannides, 2008; Rinne et al., 2008; Sorqvist et al., 2012; Woldorff & Hillyard, 1991). Those encephalic or brainstem regions upon which attention exerts its influence are probably highly task-dependent.

Adopting such analysis mechanisms additionally allows unattended sound sources to not merely be disregarded by the system, they could be processed at limited depth. Unattended representations may, for example, only be downregulated at later (cortical) processing stages, preventing their interference with the developing percept while still allowing for necessary source representation flexibility. These multi-level interactive sound representations in the brain would allow for the necessary flexibility to perform rapid adjustments of ongoing task demands, facilitation the parcellation of heavily overpopulated and highly dynamic auditory scenes (Sussman, 2017). A further interesting feature of such model is that the system does not necessarily need to have direct access to all sound features. Depending on ongoing task-demands, it can gain such input at rapid time-scales via its backward projections in a potentially reverse-hierarchical fashion.

Music Perception

The physical organization of multiple simultaneous sounds into carefully orchestrated mixtures can lead to very specific and sometimes unexpected percepts. Music composition raises such a combination of sounds to an artform, the true skill of which requires something beyond the straight-forward deployment of composition rules. Music composition has a long tradition and its heuristics are, at least in part, an indirect representation of the underlying sound analysis mechanisms of our central nervous system, more specifically auditory scene analysis.

Music essentially employs common sounds and organizes them within an elaborate framework of spectral and temporal relationships, leading to an esthetically pleasing perceptual entity, at least to most (Mas-Herrero et al., 2014; Zatorre, 2015). Testifying to this, music is in general perceived by humans as much more than the simple raw combination of its individual physical sound properties would suggest. When perceiving music, listeners employ a multitude of perceptual schemas, the makeup of which is dependent on the amount of (formal) musical training listeners have undertaken (e.g., Bregman, 1990). One such example is that prior music knowledge reduces the influence of task-irrelevant distractors during music melody recognition tasks (e.g. Bey & McAdams, 2002, 2003). Development of knowledge-based music schemas is a relevant topic in music psychology, albeit outside the scope of this thesis; an overview can be found

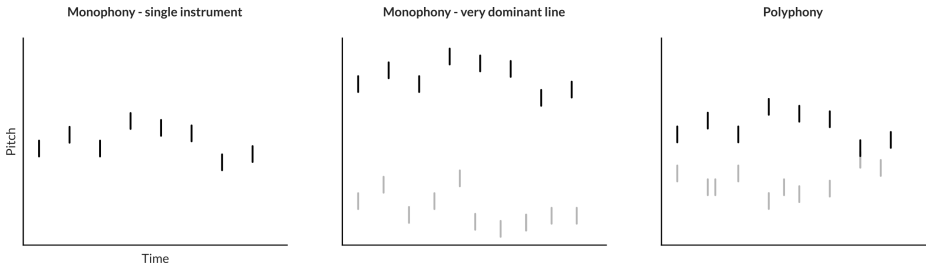


Figure 1.3: Schematic Representation of Music Structures. When a music piece is structured around a single instrument (left-most pane, black bars), or multiple instruments are present (center pane, black and gray bars) of which one melodic line is very dominant, this is typically a monophonic structure. When multiple melodic lines are present which overlap, or are in close proximity, and there is no single highly dominant line, it is typically a polyphonic structure (right-most pane).

in Deutsch (2013b).

Music typically contains combinations of horizontal and vertical relations between its elements, that is its temporal and pitch dimensions. These perceptual relationships are very much analogue to musical notation; even though the notation structures per se are not a direct analog for the physical makeup of sounds, notation does often correspond to the listeners perceived sound groupings (Bregman, 1990). Vertical relations tend to be present across multiple voices (*i.e.*, separate lines) and form larger interactive structures, leading to the possibility of perceiving merged multi-voice versions of a piece, often generated by different instruments.

Polyphony is a texture regularly found in classical music pieces and consists of two or more simultaneous lines containing independent yet proximal or overlapping melodies (Fig. 1.3). Within polyphonic music pieces, counterpoint music refers to the presence of a specific vertical dependency between the melodic lines, which creates a harmonic link between them while remaining rhythmically independent. Music pieces written in a contrapuntal structure hence possess both an appreciable horizontal structure (*i.e.*, within each instrument/voice), and an additional perceptible vertical dimension when the voices are attended to as an aggregate. By changing the physical dependencies between lines, for example their pitch-difference, it is possible to influence the degree of voice segregation or integration.

Music does not only represent an art form of interest, it can also be employed in the formal study of general auditory processing. In this context, the investigation of naturalistic ASA mechanisms could be achieved by virtue of polyphonic music. Adopting such stimuli, additionally allows for the investigation of stream integration mechanisms across instruments, as opposed to being limited to the study of classical segregative mechanisms only. Source integration is a prominent part of general ASA theory, even though its formal investigation, especially at a neurological level, has been relatively understudied in the literature (Deutsch, 2013a; Ragert et al., 2014; Sussman,

2005; Uhlig et al., 2013). Bottom-up driven changes of integration or segregation percepts can be achieved by modulating, for example, the instrument timbre or pitch differences (e.g., Bregman & Pinker, 1978; Cusack & Roberts, 2000; Deutsch, 2013a; Marozeau et al., 2013; McAdams, 2013a,b; Wessel, 1979), while top-down switches could be achieved by changing the listener's locus of attention (e.g., Besle et al., 2011; Carlyon & Cusack, 2005; Carlyon, 2003; Cusack et al., 2004; Lakatos et al., 2013; Sussman et al., 2007). Initial segregation of sufficiently physically different music voices is probably needed in order to perceive polyphony, even though the eventual percept is likely formed by integrating across the instruments under higher-level top-down influence (Bigand et al., 2000; Bregman, 1990; Gregory, 1990). Aside from the investigation of ASA mechanisms, the employment of music could provide further insight into the relationships between music-specific cognitive capacities and their extendibility to the processing of sounds in non-musical contexts.

Measuring and Analyzing Brain Activity

In this thesis, brain activity is measured with two non-invasive neuroimaging techniques, namely functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). Functional MRI provides an indirect measure of neuronal activity at the neural population level by measuring the metabolic changes in blood oxygen concentration. EEG measures the electrical activity of neural populations using electrodes placed on the scalp and intercepts the electric potential changes originating from ionic currents flowing around the neurons. To a certain degree, these two methods offer complementary insight into brain processes, where fMRI has a high spatial and a low temporal resolution, EEG has a high temporal and a low spatial resolution.

Functional MRI

Functional MRI measures changes in the ratio between oxygenated and deoxygenated hemoglobin in the blood, termed the blood-oxygen-level dependent (BOLD) hemodynamic signal (e.g., Logothetis et al., 2001). Therefore, such measures are an indirect measure highlighting where in the brain neuronal activity changes took place. Due to its good spatial resolution, this technique has been employed to investigate, among many other brain functions, the processing of sounds in the human (auditory) cortex. Classical fMRI analysis has focused on the detection of stimulus-responsive brain regions using the general linear model (GLM; Friston, 1995), which is aimed at highlighting local response differences between conditions. Response difference assessment is performed for each measured spatial location (*i.e.*, voxel) separately and is typically referred to as mass-univariate analysis. Due to the large number of voxels present in the brain, statistical assessment of significant activation changes is performed by correcting for false positives us-

ing, for example, cluster size thresholding (Forman et al., 1995) or false discovery rate correction (FDR; Genovese et al., 2002).

About two decades ago, researchers realized that fMRI activity changes related to, among others, perceptual processes can be detected by analyzing simultaneously a multitude of voxels. Haxby et al. (2001) demonstrated that image categories can be decoded based on patterns of voxels in the ventral temporal cortex, which eventually led the field to adopt pattern recognition and machine learning methods. These methods allowed investigating response pattern representations to stimulus or task modulations as opposed to focusing only on single-voxel changes, marking the beginning of multi-voxel pattern analysis (MVPA) methods in fMRI research.

One of the most common implementations of MVPA is the training of a classifier to distinguish between the experimental conditions present in the data. In the more general machine learning context, classification is defined as the problem of identifying to which data-category a previously unseen set of observations belongs. To this end, a classifier is trained based on a subset of the data (*i.e.*, training-data) containing labeled observations identifying their category membership. After completion of the fitting-procedure, the generalization of these class-identifications is assessed on the basis of independent test-data. Classifiers are part of a wider group of pattern recognition techniques, aimed at detecting regularities in complex datasets.

Even though the exact models at the heart of a given MVPA approach may vary, classification-based analysis tends to follow a specific structure diverging from GLM analysis. First, the features to be employed for classification are extracted from the fMRI data; typically response amplitude estimations at the voxel level extracted from the GLM model have been used. The number of features present in a dataset may have a large influence on model identification, especially due to the noisy nature of fMRI data, often leading to the employment of some form of feature reduction/selection. Many selection methods have been proposed, among which the more classical approach of limiting analysis to a predefined region of interest (ROI), or other restrictions based on forms of anatomical and functional combinations. Second, the (classification) model, which within neuroimaging applications tends to be of a linear nature, is fitted on the class-labeled training-data, hence linearly combining/weighing features. Third, model generalization is tested on independent testing-data, a stage during which the model assigns labels to previously unseen data. Model performance is often assessed based on accuracies, namely the proportion of correct classifications with respect to the full test-set size. Importantly, since training and testing data are generally derived from the same dataset in neuroscience, various combinations of data-splits need to be tested in a cross-validation setup and model performance is computed by averaging across all folds.

In order to assess whether classification performance is significantly above chance, its accuracies need to be compared to an empirical null-distribution estimated on the same data. To this end, the aim is to represent model performance capacities in situations where the trial labels do not match the conditions present in the data. Estimation of these non-parametric distributions in the MVPA context are mostly performed on the basis of permutations (Golland & Fischl, 2003), which fit a large number of models on the same data used for true-label classification only providing a random ordering of the class-labels. Such calculations tend to be computationally heavy, as for each cross-validation fold typically a 1000 or more permutations are performed. Accordingly, true-label classification is compared to the permutation-based accuracies distribution and the significance of above-chance classification can be assessed.

Electroencephalography

Due to the nature of the measurement's signal, EEG, as well as Magnetoencephalography (MEG), allow for the investigation of more fine-grained temporal profiles as compared to fMRI. Even though the signal is more directly related to neuronal activity, it is nonetheless an indirect measurement and mostly reflects very large amounts of spatially similarly oriented neurons firing synchronously. Since differences in electric potential are to be measured through scalp-electrodes, the signal has to be sufficiently strong to spread through both the neural and bone tissue. In order to generate an electrical field which is sufficiently strong to achieve this, it is necessary that thousands of neurons fire simultaneously at a very similar orientation, forming large aggregate ion currents flowing in the same direction and hence measurable signals at the scalp. When neurons are not similarly oriented, their currents will not sum sufficiently and hence not reach the minimum potential flow which can be measured. The classical methodology for analysis of these EEG signals, is aimed at improving its signal to noise ratio (SNR) by averaging all trials pertaining to the same experimental condition while preserving the temporal domain, resulting in an event-related potential (ERP) for each condition (e.g., Luck, 2005). Even though pattern recognition techniques have been extended into EEG analysis, the method focus of this work lies elsewhere.

Sound stream representations at the cortical level have been hypothesized to reflect to a certain degree the sound's amplitude variation over time: its envelope. A methodological paradigm has been put forward which aims at reconstructing the cortical representation of sound envelopes present in electrophysiological data, among others with EEG, MEG, and electro-corticography (ECoG; e.g., Crosse et al., 2015; Dijkstra et al., 2015; Ding & Simon, 2012; Kerlin et al., 2010; Kubanek et al., 2013; Nourski et al., 2009; O'Sullivan et al., 2015). Such analysis setup is specifically well suited for investigating the cortical representation of multiple concurrently present sounds and their potential modulation based on cognitive changes. To this end, stimulus envelopes are esti-

mated for each sound and in combination with the EEG data a sound-envelope model is fitted on the training-data. By exclusion of stimulus on and offset windows, emphasis can be placed on the influence of cognitive processes on the ongoing sound representations. After fitting, model performance is tested on the independent test-data, computing a correlation between the reconstructed and the true sound envelope of a stimulus. In order to assess model generalization for the whole stimulus set, performance is assessed in a cross-validation setup and correlations averaged over its respective folds. Model reconstruction capacities can then be compared between conditions, for example a male speaker which is attended to versus an ignored female one, and vice versa.

Outline of the Thesis

Chapter 2 introduces a novel paradigm designed for the investigation of ASA using polyphonic music, focusing on stream segregation and integration specifically. The use of a validated behavioral paradigm is of great importance in contexts where higher-level cognitive mechanisms are investigated, especially so when combined with naturalistic stimulus sets. The developed paradigm allows for the investigation of both top-down and bottom-up contributions to these mechanisms by integrating an attentive as well as an instrument timbre manipulation within the same ASA framework. Participants with limited to no musical education are the target population for investigation of these mechanisms, therefore it has been optimized for this group while still permitting for the study of these same mechanisms in highly-trained musicians. This chapter describes the paradigm and its stimuli in detail, validating its employability for the investigation of ASA in a large number of participants. Two versions of the paradigm are considered, first an attention-only manipulation, and second an attention plus timbre variant. In addition to the empirical validation, its suitability for employment in neuroimaging paradigms as well as its integration in both the music and general ASA literature is discussed. In the following chapters this paradigm is used to investigate the cortical areas involved in music scene analysis, employing different neuroscientific tools to highlight its spatial (chapter 3) and temporal (chapter 4) characteristics.

Neuronal mechanisms supporting the integration and segregation of auditory streams are investigated in chapter 3. The music paradigm is employed in combination with high-resolution fMRI at 7 Tesla and state-of-the-art analysis methodology. More specifically, it provides insight into the contribution of the temporal-frontal cortical network to the auditory scene analysis of music. Temporal-frontal network selection and consecutive multivariate classification analysis is performed in a within-subject setup, providing a powerful means of detecting subtle attentive effects at the individual subject level. Further insight into which auditory cortical areas within

the network contribute to observed effects is obtained by anatomically restricting the analysis to individually defined ROIs.

The fMRI-based investigation of chapter 3 provided for a high spatial resolution insight into music ASA processes, shedding light onto which cortical areas may contribute to its analysis, albeit at a poor temporal resolution. Chapter 4 investigates these same mechanisms, only employing a technique with a very high temporal resolution, providing valuable insight concerning effect timing as well as their possible nature. This method provides a certain level of understanding whether observed effects are based on early physically-driven (*i.e.*, bottom-up) or later top-down driven modulations. In order to achieve such classification, this chapter discusses an EEG study where the attentive modulation of stream segregation and integration is investigated by virtue of sound envelope reconstruction methods. Employment of this technique allows for the reconstruction of sound envelopes from the recorded EEG data, permitting the study whether these envelopes are modulated by the subject's locus of attention and during which delays these effects take place.

Finally, chapter 5 summarizes and discusses the findings of the empirical chapters, further reflecting upon the analysis methods employed, and outlining future research directions for both analysis methodology and empirical work.

References

Alain, C. & Bernstein, L. J. (2008). From sounds to meaning: The role of attention during auditory scene analysis. *Current Opinion in Otolaryngology and Head and Neck Surgery*, 16, 485–489.

Bee, M. A. & Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparative Psychology*, 122(3), 235–251.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403, 309–312.

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Emerson, R. G., & Schroeder, C. E. (2011). Tuning of the Human Neocortex to the Temporal Dynamics of Attended Events. *The Journal of Neuroscience*, 31(9), 3176–3185.

Bey, C. & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics*, 64(5), 844–854.

Bey, C. & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 267–279.

Bigand, E., Foret, S., & McAdams, S. (2000). Divided attention in music. *International Journal of Psychology*, 35(6), 270–278.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10, 512–528.

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 380–387.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The Perceptual Organization of Sound. Cambridge, Massachusetts: MIT Press.

Bregman, A. S. & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology: Human Perception and Performance*, 89, 244–249.

Bregman, A. S. & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1), 19–31.

Carlyon, R. P. (2003). How the brain separates sounds. *Trends in Cognitive Sciences*, 8(10), 465–471.

Carlyon, R. P. & Cusack, R. (2005). Effects of Attention on Auditory Perceptual Organization. In L. Itti, G. Rees, & J. K. Tsotos (Eds.), *Neurobiology of Attention* (pp. 317–323). Cambrige, MA: Elsevier.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 115–127.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience-Landmark*, 13(13), 148–169.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *The Journal of Neuroscience*, 35(42), 14195–14204.

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience*, 17, 641–651.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656.

Cusack, R. & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, 62(5), 1112–1120.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A*, 33(2), 185–207.

Deutsch, D. (2013a). Grouping Mechanisms in Music. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 183–248). London, UK: Elsevier.

Deutsch, D. (2013b). *The Psychology of Music*. London, UK.

Dijkstra, K. V., Brunner, P., Gunduz, A., Coon, W., Ritaccio, A. L., Farquhar, J., & Schalk, G. (2015). Identifying the attended speaker using electrocorticographic (ECoG) signals. *Brain-Computer Interfaces*, 2(4), 161–173.

Ding, N. & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1), 78–89.

Dykstra, A. R., Halgren, E., Thesen, T., Carlson, C. E., Doyle, W., Madsen, J. R., Eskandar, E. N., & Cash, S. S. (2011). Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial EEG. *Frontiers in Human Neuroscience*, 5.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009a). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, 61(2), 317–329.

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009b). Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. *PLoS Biology*, 7(6), 1–14.

Fay, R. R. (1998). Auditory stream segregation in goldfish (Carassius auratus). *Hearing Research*, 120(1-2), 69–76.

Fay, R. R. (2000). Spectral contrasts underlying auditory stream segregation in goldfish (Carassius auratus). *Journal of the Association for Research in Otolaryngology*, 1(2), 120–128.

Fishman, Y. I., Micheyl, C., & Steinschneider, M. (2012). Neural mechanisms of rhythmic masking release in monkey primary auditory cortex: implications for models of auditory scene analysis. *Journal of Neurophysiology*, 107.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, 33, 636–647.

Friston, K. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2, 189–210.

Genovese, C. R., Lazar, N. A., & Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15, 870–878.

Golland, P. & Fischl, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. In *Information Processing in Medical Imaging* (pp. 330–341). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gregory, A. H. (1990). Listening to Polyphonic Music. *Psychology of Music*, 18(2), 163–170.

Griffiths, T. D. & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5, 887–892.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293, 2425–2430.

Hulse, S. H., MacDougallShackleton, S. A., & Wisniewski, AB (1997). Auditory scene analysis by songbirds: Stream segregation of birdsong by European starlings (Sturnus vulgaris). *Journal of Comparative Psychology*, 111(1), 3–13.

Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. *Cognition*, 82(3), B113–B122.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". *The Journal of Neuroscience*, 30(2), 620–628.

Kraus, N. & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, (pp. 1–7).

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, 8(1), e53398–9.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, 77(4), 750–761.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412, 150–157.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge: MIT.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1), 43–51.

Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, 30(3), 259–274.

Mas-Herrero, E., Zatorre, R. J., Rodriguez-Fornells, A., & Marco-Pallarés, J. (2014). Dissociation between Musical and Monetary Reward Responses in Specific Musical Anhedonia. *Current Biology*, (pp. 1–6).

McAdams, S. (2013a). Musical timbre perception. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 35–68). London, UK: Elsevier Inc.

McAdams, S. (2013b). Timbre as a structuring force in music. In *ICA 2013 Montreal* (pp. 1–6).: ASA.

McAdams, S. & Bregman, A. S. (1979). Hearing Musical Streams. *Computer Music Journal*, 3(4), 26–43.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., & Courtenay Wilson, E. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research*, 229(1-2), 116–131.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biology*, 6.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., & Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *The Journal of Neuroscience*, 29(49), 15564–15574.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7), 1697–1706.

Poghosyan, V. & Ioannides, A. A. (2008). Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*, 58, 802–813.

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology*, 18.

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and Integration of Auditory Streams when Listening to Multi-Part Music. *PLoS ONE*, 9(1), 1–9.

Rinne, T., Koistinen, S., Autti, T., Alho, K., & Sams, M. (2008). Auditory selective attention modulates activation of human inferior colliculus. *Journal of Neurophysiology*, 100, 3323–3327.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2010). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3), 1–10.

Shamma, S. A. & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3), 361–366.

Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, Awareness, and the Perception of Auditory Scenes. *Frontiers in psychology*, 3, 1–17.

Sorqvist, P., Stenfeld, S., & Ronnberg, J. (2012). Working memory capacity and visual-verbal cognitive load modulate auditory-sensory gating in the brainstem: toward a unified view of attention. *Journal of Cognitive Neuroscience*, 24, 2147–2154.

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the acoustical society of America*, 117(3), 1285–14.

Sussman, E. S. (2017). Auditory Scene Analysis: An Attention Perspective. *Journal of Speech Language and Hearing Research*, 60(10), 2989–13.

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, 69(1), 136–152.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain Bases for Auditory Stimulus-Driven Figure-Ground Segregation. *The Journal of Neuroscience*, 31(1), 164–171.

Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *Neuroimage*, 77, 52–61.

van Noorden, L. (1971). Rhythmic fission as a function of tone rate. *IPO Annual Progress Report*, 6, 9.

van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America*, 61(4), 1041–1045.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, 167(3917), 392–393.

Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2), 45–52.

Winkler, I., Denham, S., Mill, R., Bohm, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R Soc. Lond B Biol. Sci.*, 367.

Winkler, I., Sussman, E., Tervaniemi, M., Horváth, J., Ritter, W., & Naatanen, R. (2003). Preattentive auditory context effects. *Cognitive, Affective, & Behavioral Neuroscience*, 3(1), 57–77.

Woldorff, M. G. & Hillyard, S. A. (1991). Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology*, 79, 170–191.

Zatorre, R. J. (2015). Musical pleasure and reward: mechanisms and dysfunction. *Annals of the New York Academy of Sciences*, 1337(1), 202–211.

"Thinking does not guarantee you will not make mistakes, but not thinking generally guarantees that you will."

Leslie Lamport

2

Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music

Based on: Disbergen, N. R., Valente, G., Formisano, E., and Zatorre, R. J. (2018). Assessing Top-Down and Bottom-Up Contributions to Auditory Stream Segregation and Integration with Polyphonic Music. *Frontiers in Neuroscience*, 12:121.

Abstract

Polyphonic music listening well exemplifies processes typically involved in daily auditory scene analysis situations, relying on an interactive interplay between bottom-up and top-down processes. Most studies investigating scene analysis have used elementary auditory scenes, however real-world scenes are far more complex. In particular, music, contrary to most other auditory scenes, can be perceived by either integrating or, under attentive control, segregating sound streams, often carried by different instruments. One of the prominent bottom-up cues contributing to multi-instrument music perception is their timbre difference. In this work, we introduce and validate a novel paradigm designed to investigate, within naturalistic musical scenes, attentive modulation as well as its interaction with bottom-up processes. Two psychophysical experiments are described, employing custom-composed two-voice polyphonic music pieces within a framework implementing a behavioral performance metric to validate listener instructions requiring either integration or segregation of scene elements. In experiment 1, the listeners' locus of attention was switched between individual instruments or the aggregate (*i.e.*, both instruments together), via a task requiring the detection of temporal modulations (*i.e.*, triplets) incorporated within or across instruments. Subjects responded post-stimulus whether triplets were present in the to-be-attended instrument(s). Experiment 2 introduced the bottom-up manipulation by adding a three-level morphing of instrument timbre distance to the attentional framework. The task was designed to be used within neuroimaging paradigms; experiment 2 was additionally validated behaviorally in the functional Magnetic Resonance Imaging (fMRI) environment. Experiment 1 subjects (N=29, non-musicians) completed the task at high levels of accuracy, showing no group differences between any experimental conditions. Nineteen listeners also participated in experiment 2, showing a main effect of instrument timbre distance, even though within attention-condition timbre-distance contrasts did not demonstrate any timbre effect. Correlation of overall scores with morph-distance effects, computed by subtracting the largest from the smallest timbre distance scores, showed an influence of general task difficulty on the timbre distance effect. Comparison of laboratory and fMRI data showed scanner noise had no adverse effect on task performance. These experimental paradigms enable to study both bottom-up and top-down contributions to auditory stream segregation and integration within psychophysical and neuroimaging experiments.

Introduction

Listening to an orchestral performance demonstrates the auditory system's extraordinary capability to both segregate and integrate sound sources within a complex mixture of simultaneously playing instruments and background sounds. While most listeners can segregate individual melodic lines, for example a flute and a harp from the mixture, the same excerpt could be differentially perceived by integrating multiple instruments into a single percept, focusing on, for example, their harmonic relationships. Mechanisms contributing to resolving such Auditory Scene Analysis (ASA) challenges have been extensively studied psychophysically and comprehensively described in Albert Bregman's (1990) work, proposing a framework for the perceptual organization of sounds. Stream segregation is responsible for parceling an auditory scene with multiple sound sources into individual acoustic events or auditory streams (Bregman, 1990; Ciocca, 2008; McAdams & Bregman, 1979; Micheyl et al., 2007). Segregation and integration of sources within mixtures of spectrally and temporally overlapping sounds is mainly driven by physical (*i.e.*, bottom-up) differences, and may be further facilitated by, among others, selective attention (*i.e.*, top-down modulations; Bregman, 1990; Brochard et al., 1999; Shamma & Micheyl, 2010). In polyphonic music, pitch and instrument timbre differences have been indicated as prominent examples of bottom-up cues (for example, Bregman & Pinker, 1978; Cusack & Roberts, 2000; Deutsch, 2013; Marozeau et al., 2013; McAdams, 2013a,b; Wessel, 1979), with top-down attention potentially modulating sound feature representation(s) or general source salience (Besle et al., 2011; Carlyon & Cusack, 2005; Carlyon, 2003; Cusack et al., 2004; Lakatos et al., 2013; Riecke et al., 2016; Sussman et al., 2007).

Polyphonic music very well exemplifies ASA in naturalistic complex auditory scenes as encountered by many listeners on a daily basis, comprising multiple sources from various instruments combined with changing degrees of spectral-temporal overlap. However, contrary to traditional cocktail-party designs, polyphonic music stimuli not only permit studying classical source segregation, they also add the possibility to investigate the relatively neglected ASA aspect of stream integration across (complex) sounds (Deutsch, 2013; Ragert et al., 2014; Sussman, 2005; Uhlig et al., 2013). Even though initial segregation of music voices is probably necessary to perceive polyphony, the simultaneous percepts of coherent melodic lines is most likely achieved by integration (Bigand et al., 2000; Bregman, 1990; Gregory, 1990), which is potentially modulated by top-down influences. A general performance benefit is observed on divided attention tasks employing polyphonic music, as compared to many experiments using other types of stimuli such as multiple simultaneous speech streams (for example, Bigand et al., 2000). Observed superior performance is hypothesized to be driven by the existence of both a perceptual and structural relationship between the multiple music voices comprising counterpoint/polyphonic

pieces. Counterpoint music contains structural relationships both across the notes of individual voices (*i.e.*, horizontal coherence) and between the individual melodic lines (*i.e.*, vertical integration). Music voices, therefore, need to have sufficient commonalities to express their musical relationship and allow their integration, by, for example, top-down processes, even though the need remains for preserving ample differentiating factors, such as pitch or timbre, to allow for their segregation.

The majority of psychophysical and neuroscientific studies on ASA have been implemented using relatively elementary auditory scenes with, for example, tones in noise or multiple alternating tone sequences (for reviews see, Alain & Bernstein, 2015; Bregman, 1990, 2015; Carlyon, 2003; Ciocca, 2008). While music listening and processing has been studied extensively in recent years (for reviews see, McDermott & Oxenham, 2008; Peretz & Zatorre, 2005; Zatorre & Zarate, 2012), very few have attempted to investigate ASA employing more complex and realistic polyphonic music (Janata et al., 2002; Ragert et al., 2014). Conversely, some studies did apply ASA segregation mechanisms to explain polyphonic/multi-part music perception (for example, Deutsch, 2010, 2013), even though no tasks have been developed to allow the study of ASA with naturalistic stimuli. The present work tries to address these factors and describes two psychophysical experiments employing polyphonic music, aimed at introducing a more ecologically valid stream segregation and integration paradigm as compared to the commonly used schematic streaming designs (for example, Bregman, 1990, 2015; Carlyon, 2003; Ciocca, 2008). We introduce a task for investigation of both stream integration and segregation with custom-composed polyphonic music stimuli, and, contrary to most previous ASA studies, provide a selective attention behavioral performance metric for both the segregation and integration of scene elements, allowing behavioral validation of task performance. The polyphonic music used was specifically designed to remain as close as possible to highly controlled stimuli typically employed in more schematic stream segregation tasks, while still being perceived as a complex music stimulus frequently encountered by listeners. Designing stimuli in this way aids task interpretation and integration within existing ASA literature, while the use of full complex music stimuli, for example extracted from existing compositions, would render the literature integration more difficult. In the current study we will introduce the task, demonstrate its validity, including its use with non-musically trained subjects, document its reliability, and make both task and stimuli available to the community for future use.

When ample physical differences between sound sources exist, such as in the case of instruments with different timbres, the integrative condition is not expected to show reduced performance compared to the segregative conditions (for example, Bregman, 1990, 2015; Moore & Gockel, 2002; van Noorden, 1977). To test this hypothesis, in experiment 1 the participant's locus of attention was varied via visual instructions. While listening to polyphonic music, partic-

ipants were asked to attend individual instruments or the aggregate (*i.e.*, both instruments) and detected rhythmic modulations incorporated within or across instruments. Our main goal was to develop a task that was challenging to non-musicians while equating difficulty across conditions, maintaining high correct scores and preserving the possibility to monitor participant's locus of attention. By reducing instrument timbre differences, more challenging segregation conditions can be created (for example, Bey & McAdams, 2003; Cusack & Roberts, 2000; Gregory, 1994; Melara & Marks, 1990; Moore & Gockel, 2002; Cusack & Roberts, 2000; Sussman, 2005; van Noorden, 1977), possibly improving performance on the less demanding task of source integration. Conversely, an increase of timbre difference could facilitate the source segregation and decrease integration performance, giving rise to a instrument-timbre interaction: more attentional resources are recruited for segregating sources with smaller physical (*i.e.*, bottom-up) differences compared to their integration, while less attentional resources are required for segregation than integration when there are larger physical differences. Such possible effects are tested in experiment 2, which adds a bottom-up manipulation to the attentive modulation framework of experiment 1 by introducing a three-level change in instrument timbre distance.

Both experiments were specifically designed to facilitate eventual investigation of top-down and bottom-up contributions to stream segregation and integration processes for complex sounds using both behavioral and neuroimaging paradigms. To generalize the task to the latter application, experiment 2 was additionally validated within the functional Magnetic Resonance Imaging (fMRI) scanner environment. Subjects were tested either in a sound attenuated chamber without background scanning sequence noise (LAB group), or during a multi-session fMRI experiment (SCAN group) with a continuous imaging sequence, allowing assessment of whether group-level task performance was affected by the addition of scanner noise to the auditory scene. The factor of background noise in fMRI scanning (Andoh et al., 2017; Belin et al., 1999; Amaro et al., 2002; Hall et al., 2014) is often ignored, but becomes of special significance in the case of stream segregation studies.

Methods

Stimuli

Twenty polyphonic counterpoint music pieces were custom-composed in close collaboration with a composer, providing desired control over acoustical content while remaining recognizable as polyphonic music. All pieces were 28 seconds (s) long and included two voices written in treble and bass clef, respectively synthesized in bassoon and cello, at tempo 60. Compositions were controlled for, among others, pitch distance between voices (never touching or

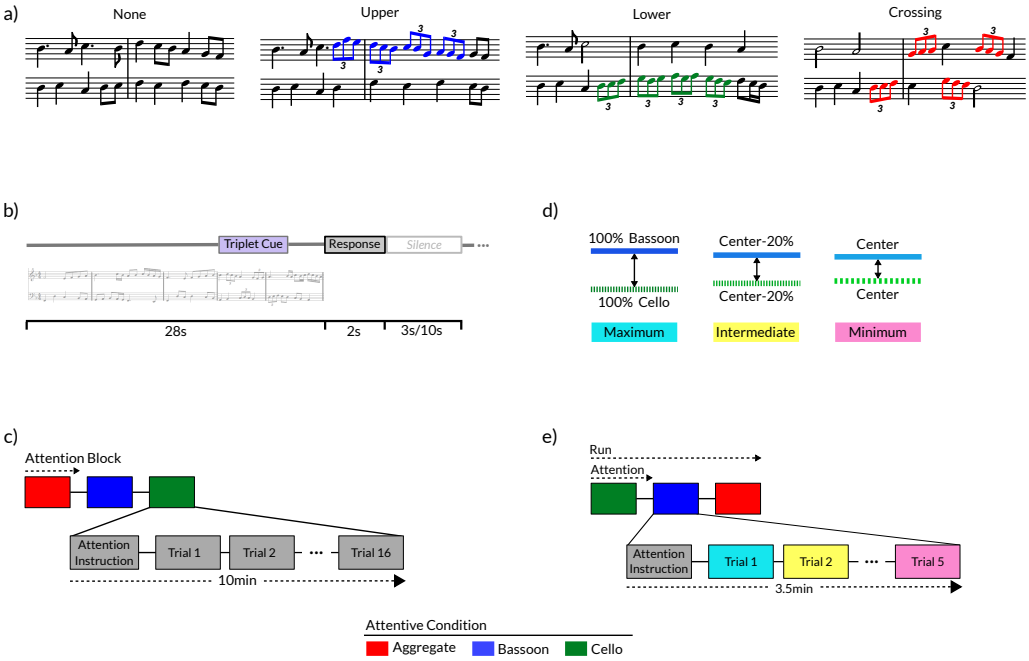


Figure 2.1: Experimental Design. Different Triplet versions for each music composition (a): no triplets, upper voice (*i.e.*, bassoon; blue notes), lower voice (*i.e.*, cello; green notes), crossing voices (red notes). Trial buildup (b) with stimulus, response window, a 3s (Experiment 1 & Experiment 2 LAB) or 10s (Experiment 2 SCAN) silence; experiment 1 (c) and 2 (e) trials presented in attentive blocks, preceded by attention instruction and silence. Experiment 2 instrument timbre manipulation (d) per triplet version: original timbre (maximum; blue), morph perception center-point (minimum; pink), or morphed 20% more towards maximum from minimum (intermediate; yellow).

crossing), rhythmic modulations, and pitch modulation size (Fig. S2.1). Tempo of 60 beats per minute was selected among other faster alternatives to allow not musically-trained individuals to detect temporal modulations within the music. Sixty-two additional unique pieces were custom-written for training and testing purposes, of which six were composed meeting the exact same requirements as the experimental pieces. All training music was unrelated to the experimental compositions and not repeated anywhere other than in their respective training round or the pre-test (see Participant selection and Training). Music pieces were synthesized from Musical Instrument Digital Interface (MIDI) files in mono for bassoon (treble clef) and cello (bass clef) independently, sampled at 44.1 kilo Hertz with a 16Bits resolution using Logic Pro 9 (Apple Inc., Cupertino, California, USA). Stimuli were combined into polyphonic pieces, Root Mean Square (RMS) equalized, and onsets-offsets exponentially ramped with 100ms rise-fall times. Stimulus processing and manipulation after sampling was performed with custom-developed MATLAB (The MathWorks Inc., Natick, Massachusetts, USA) codes.

To provide a control on the locus of selective attention, participants detected rhythmic modulations comprising a pattern of triplets incorporated in the polyphonic music. Triplets, in our spe-

cific case, are defined as three eighth notes played in the time of one beat, typically perceived by (non-musically trained) listeners as a 'speeding-up' of the music compared to its flanking notes. Temporal modulations in the form of triplets were chosen because they are orthogonal to pitch changes, facilitating detection by non-musically trained listeners and providing maximum independence from pitch-based segregation mechanisms. Patterns detected by subjects consisted of four eighth-note triplets in a row, comprising a total duration of four seconds, and were either present within bassoon (Fig. 2.1a, blue notes), or cello (Fig. 2.1a, green notes), or occurred across voices (Fig. 2.1a, red notes), or were not present. When patterns crossed voices, they started randomly with the first triplet in bassoon or cello and accordingly alternated between voices; when located within a single voice, all triplets were only present inside the respective instrument's melody. Patterns were pseudo-randomly incorporated in the second half of each excerpt between 14 and 19 seconds, and surrounding music contained no specific information concerning pattern location or presence. To prevent triplets standing out too obviously from neighboring notes, the patterns followed the excerpt's melody. Adopting such a design resulted in stimuli which only differed as to the inclusion and position of triplets. Incorporating the same four-triplet pattern within and across voices ensured participants detected the same pattern independently of condition. Experimental stimuli are available for download via the Zatorre lab's website[*].

Instrument timbre was manipulated for each melody separately via interpolation using the STRAIGHT[†] vocoder speech manipulation software tool in MATLAB (Kawahara & Matsui, 2003). Morphing was performed individually for each melody (*i.e.*, voice), interpolating the melody from the version played by a bassoon towards that played by the cello, hence modulating the melody's timbre/instrument. To achieve this, time-frequency landmarks were created on each instrument's synthesized melody, with landmark time centered on the middle of each note and frequency tagged as the note's f0. Within each music voice, instrument timbres were morphed by logarithmic interpolation of spectral density and aperiodicity. For a subset of five experiment 1 compositions (1, 4, 5, 8, and 10), instrument timbre was manipulated by morphing individual voices played by their original instrument towards the other instrument in 10% increments, for example leading to the following combinations: 100% bassoon & 0% cello, 90% bassoon & 10% cello, 80% bassoon & 20% cello, *et cetera*. Resulting stimuli were combined into polyphonic pieces, RMS equalized, exponentially ramped with 100ms rise-fall times, and filtered per individual channel with Sensimetrics equalization filters in MATLAB.

[*] http://www.zlab.mcgill.ca
[†] http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/

Participants

Twenty-nine adult volunteers (18 women; age 23.4 +/- 3.7 years, *mean +/- standard deviation*) with self-reported normal hearing, motor, and vision abilities participated in experiment 1. None of the participants spoke a tonal language and all had less than two years of (formal) music training on a lifetime basis with instruments other than bassoon or cello, as assessed via the Montreal Music History Questionnaire (Coffey et al., 2011). After completing experiment 1, a group of 19 listeners also participated in experiment 2 (13 females, age 22.7 +/- 2.6 years). Participants in experiment 2 were initially recruited for a fMRI experiment, however, when not eligible for MRI due to safety concerns, the experiment was completed in a sound-attenuated chamber (N=9; LAB). The remaining subjects (N=10; SCAN) performed the task inside the fMRI scanner over the course of two or three scanning-sessions. Subject assignment to LAB or SCAN group was solely depended on their MRI eligibility, and hence should not have resulted in any bias with respect to task performance ability. Volunteers were mostly students recruited from McGill University (N=7) or Maastricht University (N=22), with both McGill and Maastricht volunteers participating in experiment 1, and a subgroup of Maastricht volunteers in experiment 2. Subjects provided written informed consent and experimental procedures were approved by the ethical committees of each university.

Experiment 1

Task

Participants completed a forced-choice delayed-response target detection task within or across music voices, attending to the same instrument(s) during a block of 16 trials (Fig. 2.1c); before initiation of each block they were visually instructed to attend to the bassoon, cello, or aggregate. After each stimulus ended, listeners indicated via a button press whether the triplet pattern was present in the instrument(s) instructed to be attended or not. Post-stimulus responses were adopted to reduce the influence of cognitive decision and motor processes in the stimulus presentation window, specifically beneficial for task employment in neuroimaging experiments. Depending on the attentive condition, triplet presence and/or position differed: in the attend to both instruments condition, half of the trials had triplets crossing over the voices and half contained no triplets; in the attend to bassoon condition, half the trials contained a target pattern in the bassoon voice, as a control one-fourth of the trials contained triplets in the unattended (cello) voice, and one-fourth of the trials contained no triplets; similarly, in the attend cello condition half the trials comprised triplets in cello, one-fourth in bassoon and one-fourth no triplets. Equal distribution between target and no-target trials was adopted to prevent response bias. No-triplet trials along with triplet patterns in the opposite voice were employed

to check for false alarms and whether subjects switched attention between instruments as opposed to exclusively focusing on a single voice. Furthermore, opposite-voice patterns provided information if the correct instrument was attended and, in case of a large number of misses, allowed determining whether triplet detection in the opposite voice was achieved. Even though it is not possible to determine with absolute certainty whether a subject was attending to the cued instrument(s) in the various conditions, high performance on the task in conjunction with the above-mentioned measures does provide a strong indicator of a subject's capacity to segregate and integrate music streams.

Trial duration was 33s and comprised a stimulus of 28s, response window of 2s, and a 3s silence (Fig. 2.1b). After nine trials simulating the experiment, 96 trials from 16 (N=7 participants; McGill University volunteers), or a sub-set of 10 (N=22; Maastricht University volunteers), unique compositions were presented over six attentive blocks. Each stimulus block represented a condition and all conditions were repeated twice, each time with a unique stimulus order. Within a participant the same conditions could not follow each other and the order within the first and second block of three conditions had to be unique (e.g., ACB-CAB); condition ordering was balanced across subjects. Stimulus presentation order was pseudo-random, controlling that within a block of sixteen trials, stimuli of thesame composition could not follow one another, and a composition could not be repeated more than once. Stimuli were delivered through Shure SRH1440 professional open-back headphones (impedance 37Ω; Shure Inc., Niles, Illinois, USA) at approximately 85 dB SPL in a sound-attenuated chamber via a Creative Sound Blaster Audigy 2ZS (Creative Technology Ltd., Singapore) sound card, employing Presentation 17.0 (Neurobehavioral Systems Inc., Albany, California, USA) for stimulus presentation and response recording. Before participation in the experiment, subjects completed a training session and a pre-test for learning assessment; see Participant selection and Training for details.

Experiment 2

Task

In experiment 2, a manipulation of bottom-up information (timbre) was added to the experiment 1 design, keeping all task and stimulus aspects not otherwise mentioned below equal to experiment 1. The difference between instrument timbres was varied across three discrete levels while subjects performed the attentive task as described in experiment 1. Timbre morphs were combined to create three instrument timbre distances: each melody played by their respective original instruments (*i.e.*, no timbre manipulation; maximum; Fig. 2.1d, blue), minimum timbre difference between instruments (minimum; Fig. 2.1d, pink), and 20% closer towards maximum from the minimum distance values (intermediate; Fig. 2.1d, yellow). Minimum timbre distance

between voices was determined perceptually for each subject in a separate experiment, rating their instrument perception for all timbre morphs per voice (see Timbre Perception). Individual matching of timbre distance was adopted to account for subject variation with respect to their perceptual center points, based on a pilot experiment suggesting such differences.

Participants attended the same instrument(s) during an attentive block of five trials, each comprising a stimulus of 28s, response window of 2s, and a 3s (LAB) or a 10s (SCAN) post-silence (Fig. 2.1e). After several practice trials, a total of 90 (LAB) or 135 (SCAN) trials were presented across six (LAB) or nine (SCAN) runs of three attention blocks each (Fig. 2.1b). Composition version distribution across conditions was equal to experiment 1, with the only exception that eight target trials and seven control trials were included in a run. Within the bassoon and cello conditions, this distribution resulted in control trials comprising uneven numbers of no-triplet versions and opposite-to-attention-voice triplet versions (three or four of each). To mitigate any effects of such imbalance, their numbers alternated across experiment repetitions: three-four or four-three. Stimulus order was pseudo-random, controlling that each timbre distance version of each composition was covered by all conditions over the course of three runs/nine attentive blocks. Within attention blocks, compositions could occur only once, same timbre distances could not follow, a composition's timbre distance had to occur at least once, and the same timbre distance could not occur more than twice. Within a run, stimuli could only occur once, the first stimulus of a consecutive attentive block could not be the same as last of the previous, and number of timbre distance occurrences for each composition had to be equal. In one experiment repetition (*i.e.*, three runs/nine attentive blocks), all conditions uniquely occurred at each position within the three-block sequence, for example: ABC-BCA-CAB. For all three experiment repetitions (*i.e.*, nine runs/twenty-seven attentive blocks), condition order blocks appeared in all positions, for example: repetition 1) ABC-BCA-CAB, 2) BCA-CAB-ABC, and 3) CAB-ABC-BCA. Across participants, condition run order was balanced, for example: participant 1) BCA-CAB-ABC versus 2) CBA-BAC-ACB. Stimuli were presented through Sensimetrics (Sensimetrics Corporation, Malden, Massachusetts, USA) S14 ear-buds at approximately 83 dB SPL via a Creative Sound Blaster Audigy 2ZS sound card (LAB). During the fMRI sessions (SCAN) stimuli were presented via Sensimetrics S14 ear-buds and a Creative Sound Blaster X-Fi Xtreme Audio sound card at around 94 dB SPL, a gain of approximately 30dB over the scanner sequence noise.

Timbre Perception

During a separate session, each participant's minimum instrument timbre distance point was perceptually determined by rating their perception of all timbre morphing steps per individual voice. Each subject was presented with all 10% morph steps to allow assessment of their changing percept, as we were not aware of any other data quantifying these timbre modification ef-

ID	Triplet Type	Melody	Instrument(s)	Triplet Location	Task (test-stim)	Min. Correct (%)
R1	Single	Scales	Bassoon	Single Voice	PA (6)	100
R2			Cello			
R3		Basic	Bassoon			
R4			Cello			
R5		Complex	Bassoon			
R6			Cello			
R7	Pattern		Bassoon & Cello	Bassoon	PA (5)	85
R8				Cello		
R9				Crossing		
R10		Experimental		Variable	PA (8)	
Pre-test					Exp (24)	

Table 2.1: Training Structure. Training rounds and pre-test with their respective triplet types, melody complexity, included instruments, triplet location, task and number of test stimuli, and minimum score needed to pass the round. R = training round, PA = rate triplet(s) as present or absent, Exp = experimental task.

fects. Listeners were first habituated to the unaltered instrument timbre, corresponding to maximum timbre difference across voices, with 10 stimuli per instrument. Next, we assessed with 10 different test stimuli per instrument whether they correctly identified the timbre of both bassoon and cello. Subjects ranked their timbre morph perception of 176 trials on a one-to-five scale: 1=bassoon, 3=intermediate, 5=cello. Morphs were pseudo-randomly presented over eight equal-length blocks, controlling that the same composition or the same voice did not follow each other. The rating scale was visually presented throughout the experiment, fading to the background during stimulus presentation and turning brighter for the two-second response window which was followed by a two-second silence. Stimuli spanned all 10% timbre morphing steps, giving a total of 11 versions per instrument voice per composition. Perceptual midpoints were determined by fitting a sigmoid to all voice's data points r, with minimum a=1 (bassoon percept rating) and maximum b=5 (cello perception rating). Center point $x50$ as well as slope m was estimated per individual voice by nonlinear regression with iterative least squares estimation and initial values m=1 and $x50$=voice mean rating:

$$S = 1 + \frac{b - a}{1 + 10^{(x50 - r) * m}}$$

Sigmoid center point $x50$ was subsequently selected as the voice's timbre perception center, and morphs were combined into the three instrument timbre distances (see Experiment 2 Task): maximum (Fig. 2.1d, blue, i.e. original instruments), minimum (Fig. 2.1d, pink, *i.e.*, perceptual center for each instrument), and intermediate (Fig. 2.1d, yellow, *i.e.*, perceptual center minus 20 percent).

Participant Selection and Training

Participants received experiment 1-specific training, exposing them over 10 training-rounds to music of increasing complexity, ranging from scales including isolated triplets in a single voice until polyphonic melodies at melodic complexity equal to experimental stimuli and including the triplet pattern (Table 1). Training rounds consisted of initial instructions including examples and, to assess learning, several test stimuli with varying performance requirements (see Table 1). Protocols provided the option to repeat examples as desired and were developed to be self-explanatory as well as adaptive to participant performance. Feedback was presented after each trial requiring response, and if training round test performance was insufficient, the full round could be repeated maximum twice. After training, generalization was tested via a pre-test simulating a shortened version of experiment 1, employing four custom written compositions across 24 trials. Training was completed in two different groups, the first (N=14) started with training and completed experiment 1 if requirements were met (N=10). The second group (N=68) first completed the perceptual rating experiment of instrument timbres and if perceiving differences between morphs (N=21), continued to the training phase, which was completed successfully by the 19 subjects who participated in experiment 2. Pooling over both groups, 83 percent of participants could be successfully trained.

Analysis

Responses to both experiments were classified as hits H, misses M, false alarms FA, and correct rejections CR. Due to scores occurring close to ceiling, d-prime values were edge-corrected (see for example, Stanislaw & Todorov, 1999) for visualization, Bayesian model initiation (see below), and comparison purposes only:

$$d'_{ec} = \Phi^{-1}\left(\frac{H + 0.5}{H + M + 1}\right) - \Phi^{-1}\left(\frac{FA + 0.5}{FA + CR + 1}\right)$$

where Φ^{-1} denote the inverse normal cumulative distribution function with $\mu = 0$ and $\sigma = 1$.

Differences between experimental conditions were statistically evaluated with an ANOVA-like hierarchical Bayesian model including multiple grouping variables, with each subject contributing

measures to all groups. Hierarchical models are particularly well suited to describe data from individuals within groups, comprising parameters for each individual as well as higher-level group distributions, allowing integrating group and individual parameters in the same model. Such an integration has several advantages, among which the possibility to correctly estimate the variance due to subject effects with different/smaller sample sizes across groups. Additionally, these models can seamlessly handle participant performance close to ceiling, an unequal number of trials per condition, and possible heteroscedastic variances across conditions (based on a pilot experiment suggesting such variance differences). One of the strengths of incorporating Bayesian methods in a hierarchical framework, comprises the possibility of reallocating the model's parameter value credibility over more restrictive options when more data is added to the model, providing as output a distribution of credible parameter values (e.g., a condition's effect) which inherently capture the estimated parameter's uncertainty. Furthermore, the use of Gibbs sampling in the Bayesian framework allows to perform inference on those models which cannot be analytically derived, as is the case for our model, preventing the need for approximations/simplifications to make the problem tractable. Estimation and inference with hierarchical models presents many challenges in the standard (frequentist) setting, therefore, in this work, both for computational reasons and model flexibility, we chose to employ instead Bayesian estimation.

The use of a standard frequentist analysis on d-prime would be suboptimal in our case, due to several subjects performing at ceiling. It would be necessary to perform such analysis on edge-corrected d-prime values, since ceiling d-prime is infinite, which, among others, leads to a loss of sensitivity in the higher ranges. Furthermore, submitting an estimated d-prime to a standard statistical test, for example an ANOVA, would rely on the assumption that all measurements have similar uncertainty. However, this is not the case for d-prime values close to ceiling, which have a different variance compared to those in the lower ranges. The most employed statistical tool to handle such heteroscedasticity is a hierarchical (*i.e.*, mixed effects) model, which can weigh different measurements based on their uncertainty and produce a reliable population estimate. These models can be estimated in a frequentist or a Bayesian framework, given the model considered in our work the Bayesian approach was most suited. Bayesian hierarchical models provide further advantages over both frequentist and non-hierarchical models, however full coverage of strengths, weaknesses, and differences between models is beyond the scope of this paper; for an introduction to Bayesian data analysis, see Kruschke (2014).

Hierarchical model parameter estimate ranges will be expressed as Highest Density Intervals (HDIs; for example, Kruschke, 2014), whose range spans x-percent of the parameter estimation distribution and is analogous to the parametric confidence interval; an HDI of, for example, 95% from a standard normal distribution would extend from -1.96 to 1.96. When an effect's HDI

range includes zero it indicates there is probably no effect of respective condition. Models were estimated with JAGS[‡] (Just Another Gibbs Sampler, version 3.3.0) via its Matlab integration MATJAGS[§] (version 1.3.1) employing Gibbs sampling Markov Chain Monte Carlo (MCMC) simulations. JAGS models are defined by nodes, which in the model definition are written as either a stochastic relationship "$\sim$" (*i.e.*, random variable), or a deterministic relation "$\leftarrow$" where the respective node value is determined by its parents. Each model was estimated with 10.000 MCMC samples and a burn-in of 2.000 samples. Edge-corrected d-primes were used as model initiation variables, providing condition specific values for each subject from which to start model fitting.

Correlation of subject performance between experiment 1 and experiment 2 for those subjects who completed both experiments (N=19) was performed using Spearman linear rank correlation on correct rates over all trials for experiment 1 and maximum morph distance only trials for experiment 2. Experiment 2 correlation between correct rates for all trials and subject morph effect, was calculated with a Spearman linear rank correlation by subtracting accuracy for all minimum morphing distance trials from all maximum distance trials. Correct rates were selected due to a different number of trials between comparisons, hence a difference in their maximum edge-corrected d-prime values.

Hierarchical Model Specification

Experiment 1 model is a simplification of the experiment 2 model (Fig. S2.2), hence only the latter is discussed here in detail. Subject observations per condition were labeled as measurements m, three attention and three timbre distance conditions resulted in nine samples per participant and a total of m=171. Hits (H) and false alarms (FA) were, respectively, binomially modeled with the number of hits n_{hit}, trials including triplets n_{tri}, false alarms n_{fa}, and trials excluding triplets n_{ntri}; note that this modeling of d-prime is not equivalent to the edge-correction described above. Hit and FA distributions were transformed using the standard normal cumulative distribution function Φ, within which the bias-model ($bias_m$) links to the d-prime model. By modeling H and FA with the Binomial distribution, we assume all trials are independent and identically distributed:

$$H \sim Bin(n_{hit}, n_{tri})$$
$$a_{hit} \leftarrow \Phi(0.5 * d'_m - bias_m)$$
$$FA \sim Bin(n_{fa}, n_{ntri})$$
$$a_{fa} \leftarrow \Phi(-0.5 * d'_m - bias_m)$$

[‡]http://mcmc-jags.sourceforge.net
[§]http://psiexp.ss.uci.edu/research/programs_data/jags/

Group d-prime values over measurements m were modeled as normally distributed with mean $\mu_{d',m}$ and variance $\sigma_{d',attn}$, by which we assume equal variance across subjects and timbre conditions but not for attentive conditions:

$$d'_m \sim \mathcal{N}(\mu_{d',m}, \sigma^2_{d',attn})$$
$$\sigma_{d',attn} \sim Unif(0, 4)$$

Expected values of the d-prime distribution were estimated with an ANOVA-like model, employing normally distributed parameters: intercept β_0, attention β_{att}, timbre β_{timb}, attention-by-timbre interaction β_{a*ti}, and subject β_{subj}:

$$\mu_{d',m} = \beta_{0,m} + \beta_{att,m} * x_m + \beta_{timb,m} * x_m + \beta_{a*ti,m} * x_m + \beta_{subj,m} * x_m$$
$$\beta_{0,m} \sim \mathcal{N}(0, 1000)$$
$$\beta_{att,m} \sim \mathcal{N}(0, \sigma_{\beta_1^2}); \sigma_{\beta_1} \sim \Gamma(1.64, 0.32)$$
$$\beta_{timb,m} \sim \mathcal{N}(0, \sigma_{\beta_2^2}); \sigma_{\beta_2} \sim \Gamma(1.64, 0.32)$$
$$\beta_{a*ti,m} \sim \mathcal{N}(0, \sigma_{\beta_3^2}); \sigma_{\beta_3} \sim \Gamma(1.64, 0.32)$$
$$\beta_{subj,m} \sim \mathcal{N}(0, \sigma_{\beta_4^2}); \sigma_{\beta_4} \sim \Gamma(1.64, 0.32)$$

where $\Gamma(a, b)$ denotes a gamma distribution with shape a and rate b.

Distribution of $\beta_{0,m}$ was centered on zero with large variance to allow a wide range of intercept values, capturing possible d-prime differences between conditions. Beta prior values for all other predictors are limited in their range via variance σ_{β_n}. Allowing Beta standard deviations to range freely between close to minimum and maximum probability would result in an unrealistically large standard deviation and range. Such an exaggeration would have too large an influence on our data set with only a moderate number of samples. Based on suggestions by Kruschke (2014, Chapter 21) and considerations concerning number of trials as well as d-prime range, gamma prior values were restricted to $\Gamma(1.64, 0.32)$, which has mode 2 and standard deviation 4. Experiment biases were modeled equivalently to the described d-prime model, however, for conciseness, only the d-prime model is described in detail, even though the bias is an integral part of the full hierarchical model (see Fig. S2.2).

Experiment 1 resulted in three measurements per participant, a total of m=87. Hierarchical model implementation was identical to experiment 2, with only exception being a parameter reduction

on $\mu_{d',m}$ to intercept, attention, and subject effects:

$$\mu_{d',m} = \beta_{0,m} + \beta_{att,m} * x_m + \beta_{subj,m} * x_m$$

Results

Experiment 1

Inspection of Figure 2a suggests that there was no difference between attention conditions for edge-corrected d-primes in experiment 1. Bayesian hierarchical model contrasts between all pairs (Fig. 2.2f-h) confirmed that none of the attention effects differed from the grand mean (*i.e.*, subject factor). The grand mean demonstrated that mean performance was relatively comparable across subjects (95% HDI = [3.14 3.76]; Fig. 2.2b); the bias model terms for each condition were strongly centered on zero. Results indicate that the training was successful and subjects were able to complete the task with overall high correct scores, while at the group level task difficulty did not differ across attentive conditions. No systematic group difference was found for False Alarms generated by control trials containing no triplets versus those with triplets in the unattended instrument. Assessment of accuracy across all trials for included compositions showed comparable values at the group level (Fig. S2.1c). Split-half reliability on edge-corrected d-prime values for the full experiment yielded a correlation coefficient of 0.74 ($p<.001$).

Experiment 2

Subject timbre perception center points for the upper voice were determined at 0.4 (*i.e.*, 60% bassoon & 40% cello; N=1), 0.5 (N=11), 0.6 (N=6), and 0.7 (N=1); for the lower voice 0.4 (N=3), 0.5 (N=10), and 0.6 (N=6). Group mean values of edge-corrected d-prime (Fig. 2.3) indicated that intermediate and minimum timbre distance for both bassoon and cello conditions resulted in lower d-prime values compared to their respective maximum timbre distances as well as all three aggregate timbre distances. The observed change in edge-corrected d-prime values was caused by both an increase in False Alarm rate and a decrease in Hit rate (Fig. S2.4), confirming effects were not driven by participant bias; no False Alarm bias was found across subjects for trials with no triplets versus unattended instruments. Accuracy levels across all trials for included compositions were comparable at group level (Fig. S2.1d). Edge corrected d-prime scores computed over all trials resulted in a split-half correlation of 0.84 ($p<.001$), and subject performance on experiment 2 was predictable by experiment 1 scores ($r=0.49$, $p=0.031$; Fig. 2.5a).

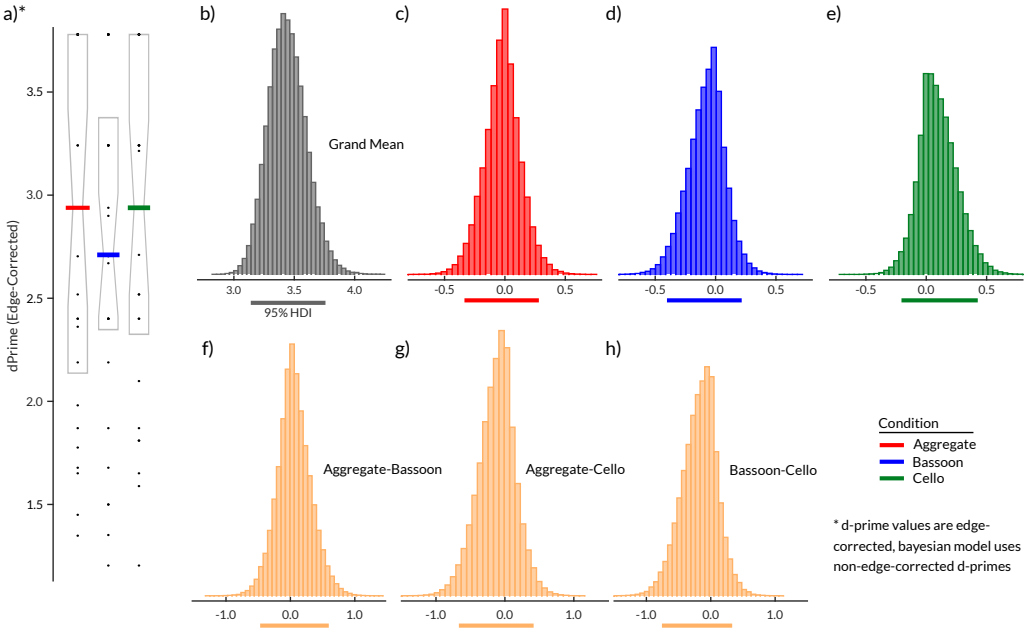


Figure 2.2: Experiment 1 Results. Group data (N=29) for edge corrected d-prime per condition (a), with median (horizontal bar) and 25th-75th percentile. Bayesian hierarchical model (b-h), with group grand mean (b), attention effects (c-e), and contrasts between attention conditions (f-h). Horizontal lines below b-h indicate 95% HDI.

Inspection of the Bayesian model grand mean ([4.02 5.34]; Fig. 2.4a), demonstrated a moderate mean performance variation across subjects. Attention main effect did not differ from the grand mean (Fig. 2.4b), nor did subsequent attentive condition contrasts (Fig. S2.3) indicate a main effect of attention. Timbre distance effect (Fig. 2.4c) of the minimum timbre difference condition did differ from zero ([-1.18 -0.24]), indicating that this condition may be more difficult than both the maximum and intermediate timbre distances. Respective contrasts confirmed that the timbre distance main effect was driven by the minimum timbre distance, with both maximum-minimum ([0.25 1.87]; Fig. 2.4g) and minimum-intermediate ([0.27 1.88]; Fig. 2.4h) differing from zero. The within-attention condition timbre distance effects (Fig. 2.4d-f) and their respective contrasts (Fig. S2.3), yielded no differences between timbre distance conditions. This observation confirmed that within attention condition timbre distances did not differ in difficulty; the model bias terms were centered on zero and showed no effect of condition.

Even though no timbre distance effect was found within attention conditions, the data displayed a trend towards the minimum timbre difference being more difficult than both maximum and intermediate timbre distances when segregating (bassoon and cello conditions), while the opposite was observed when integrating (aggregate condition). Further inspection of several contrasts testing for interaction effects did not indicate any differences (Fig. S2.3). However, when

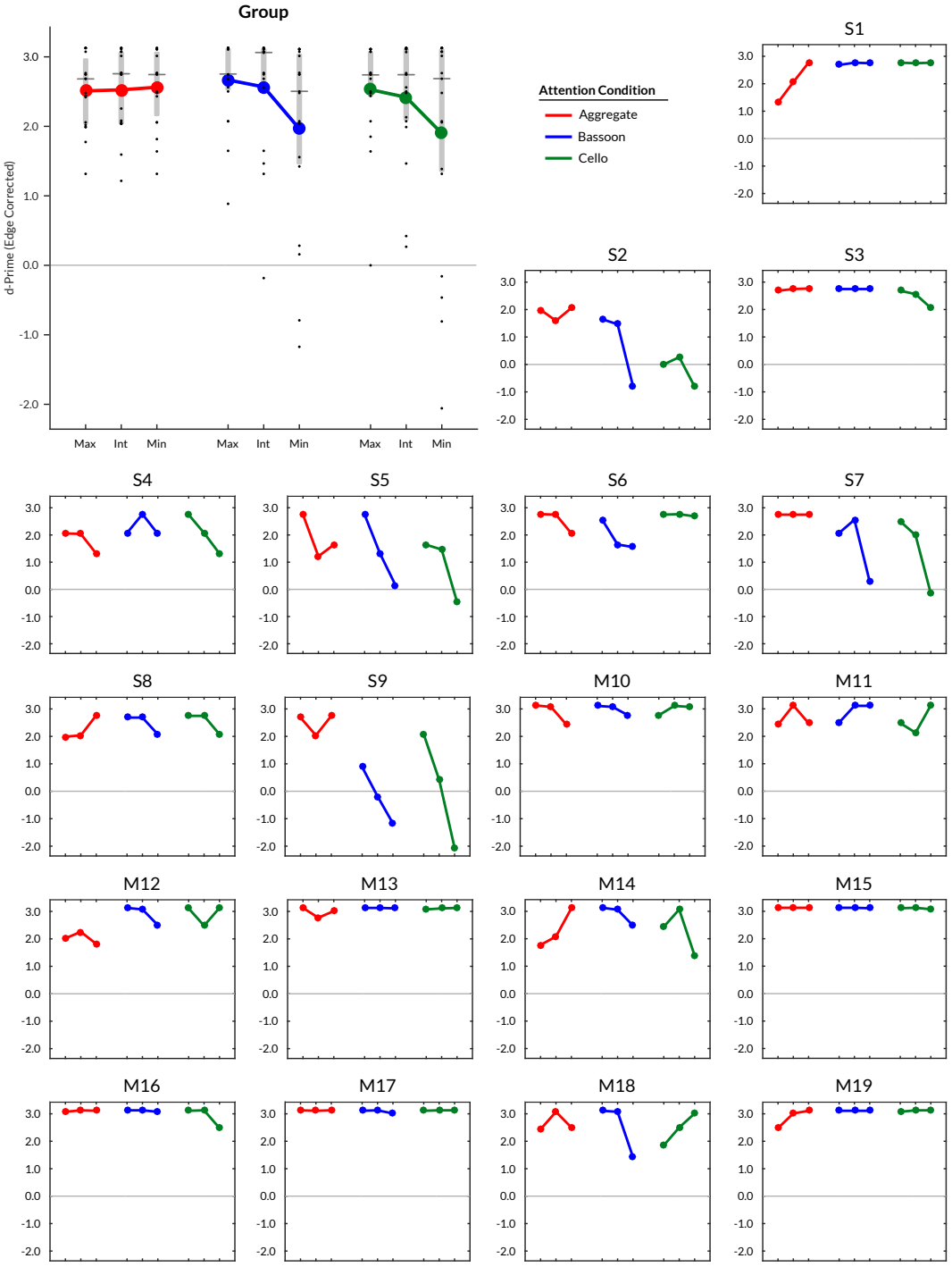


Figure 2.3: Experiment 2 Behavioral Data. Group and individual edge corrected d-prime results (N=19) per condition and instrument morph distance. Timbre distance: Max = maximum, Int = intermediate, Min = minimum. Vertical gray bars in group plot range 25th-75th percentile; horizontal gray lines median. Participants: S = experiment in sound attenuated chamber (LAB), M = experiment during fMRI (SCAN).

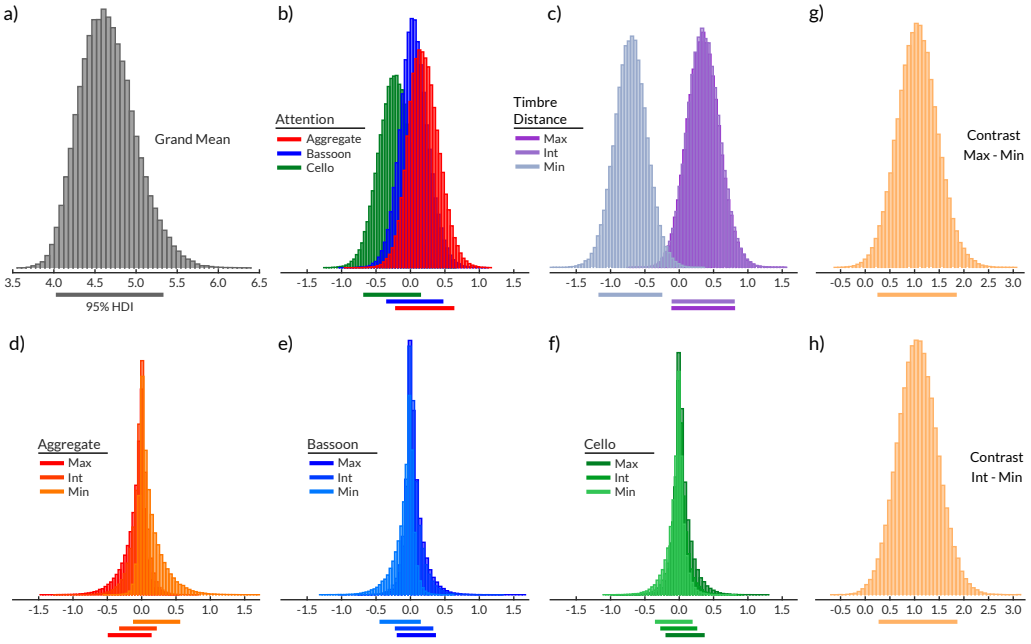


Figure 2.4: Experiment 2 Model Results. Bayesian hierarchical model grand mean (a), attention effects (b), morphing effects (c), morphing effects per attentive condition (d-f), all contrasts with a 95% HDI differing from zero (g-h; see also Fig. S2.3). Timbre distance: Max = maximum, Int = intermediate, Min = minimum. Horizontal lines below histograms, 95% HDI.

correlating subject morph effects, computed by subtracting maximum and minimum timbre distance scores, with their correct rates over all trials, those with lower overall scores showed the largest influence of morphing distance ($r=.76$, $p<.001$; Fig. 2.5b), suggesting that behavioral effects of timbre distance may be masked in those subjects performing close to or at ceiling.

Comparing experiment 2 LAB (.90 [.73 .95], *Median [Inter Quartile Range]*) and SCAN (.99 [.94 1.0]) group data for correct rates over all trials, the addition of scanner noise to the scene did not have a detrimental effect on task performance; experiment 1 LAB (.90 [.83 .96]) and SCAN (.92 [.91 .97]) group performance did not show differences.

Discussion

We presented a novel paradigm developed to investigate both top-down and bottom-up modulations of auditory stream segregation and integration with custom-composed polyphonic music suitable for use by musically untrained listeners, and adaptable to neuroimaging protocols. In experiment 1, subjects listened to two-part polyphonic music containing triplet patterns that served as attentional targets, and were instructed to attend to individual instruments (segregation), or to the aggregate (integration). Experiment 2 added a bottom-up modulation of instru-

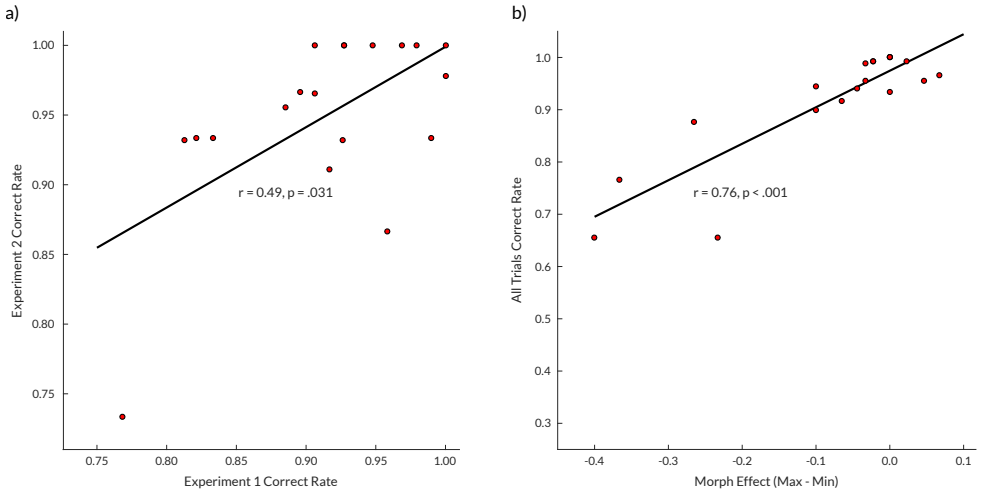


Figure 2.5: Experiment and Trial Correlations. (a) Experiment 1 and experiment 2 (maximum timbre distance trials only) correlation between correct rates across trials; for experiment 1 only those subjects who completed both experiments were included (N=19). (b) Correlation between experiment 2 subject morph effects and correct rates across all trials. Morph effect calculated by subtracting correct rates over all minimum morphing distance trials from correct rates of all maximum distance trials.

ment timbre distance into the attention-modulation framework of experiment 1. Analysis of both experiment 1 and experiment 2 indicated that listeners were able to correctly identify the target in both tasks, after only modest training, at high performance levels. We observed no group-level performance difference between attentive conditions or instrument timbre distances, for both integration and segregation. In a subset of subjects, however, there appears to be a trend towards smaller timbre distances leading to a performance decrease, more specifically among those participants showing overall lower performance (Fig. 2.5b), even though no significant interactions were found (Fig. S2.3). As demonstrated in experiment 2, the addition of scanner noise to the task had no adverse effect on task performance, validating the task's suitability for fMRI studies with continuous pulse sequences. Aside from performance metrics, subjects tested in the scanner indicated having no difficulty segregating stimulus from scanner noise due to both the loudness difference and its continuous/repetitive nature. Overall, the findings demonstrated that non-musicians could be trained to both detect triplet patterns and reliably switch attention between scene elements, enabling the task to be employed in experiments studying stream segregation and integration in a natural listening context.

Task Considerations and Future Applications

High correct scores within our paradigm are desirable, both from the perspective of task compliance and task suitability for imaging experiments. We believe that a subject's capacity to detect the triplets correctly in individual voices, or across voices in the integrated condition, provides a

strong indication they were either segregating or integrating, respectively. Each occurrence of triplets was incorporated into the melodic structure so as not to stand out from the surrounding music; this feature was explicitly designed during the composition in order to prevent any form of triplet pop-out, as this could lead to unintentional attention-target switches or alternative task strategies. The experimental design specifically focused on creating stimuli and task conditions which require listening effort, but are nonetheless feasible, to ensure that subjects are engaged in performing the desired task. Subjects reported that they needed strong attentional engagement to perform the task, both inside and outside of the scanner environment, especially due to their limited musical training. The subjects' capacity to detect the triplets correctly is primary evidence that they managed to segregate the mixture into individual streams. If listeners had not streamed the two melodies, they would not have been able to correctly respond whether the triplets were present or absent within a single instrumental voice. If the two streams are not segregated, the main remaining source of information differing between them would be rhythmic cues (tone onsets and offsets), based on which it would not be possible to assign the triplet's occurrence to one or the other of the voices. What we aim to investigate with this design is the brain's mechanisms which allows a listener to experience distinct melodic voices (or integrate across them) despite that the input arriving at the ear consists of a single mixed waveform of all sounds present in the scene.

Our paradigm was not designed for the detection of subject or condition differences, as demonstrated by a partial ceiling effect on the scores. Future iterations of the experimental protocol can possibly be sensitized to these effects by making the task more difficult for high-performing subjects. Pitch distance between melodic lines could, for example, be parametrically varied at an individual level to determine minimal pitch disparity needed for segregation (for examples see, van Noorden, 1977; Bregman, 1990) and hence maximal engagement of top-down processes. Conversely, pitch differences could be increased until segregation becomes almost inevitable for maximum reliance on bottom-up processes. Such designs would allow studying stream segregation and integration at various rates of top-down and bottom- up reliance. Increasing difficulty without adjustment of pitch could also be achieved by further reduction of timbre distance between instruments (for examples see, Cusack & Roberts, 2000; Bregman, 1990). As mentioned, timbre distance reduction appears to have a more pronounced effect on subjects who are not performing close to ceiling, compared to those who are (see Fig. 2.3 & Fig. 2.5b). This observation could be explained by a reduced cognitive load in high-performing subjects, allowing for compensation of the difficulty increase caused by timbre distance reduction, and possibly preventing emergence of a within-condition timbre effect and the hypothesized interaction effect. Insight into whether this hypothesis holds could be provided by testing these subjects with a timbre distance smaller than that based on their perceptual center points; alter-

natively, a higher tempo of the music could be adopted to make the task more challenging.

Due to the observed general performance increase/learning effect between experiment 1 and experiment 2 (Fig. 2.5a), for future use we recommend that the two tasks be carried out in separate groups, possibly providing more challenging conditions and leading to a timbre-distance sensitivity of the performance metric. Subsequent task iterations could adopt the triplet presence or absence response immediately upon target detection, allowing, among others, the investigation of possible reaction time differences. Measuring reaction time could provide a handle on both intra- and inter-subject difficulty differences between conditions or trials, highlight possible cognitive load differences between subjects, and allow further investigation of whether specific stimuli are driving, or inhibiting, factors. A delayed response was adopted in the current setup to allow testing for task applicability to the neuroimaging setting, in which such a design is favored to reduce signal contamination by motor and decisional components. Even though the current sample showed better performance in the group tested in the scanner compared to the lab, we do not believe that subject performance is likely to be enhanced by being tested inside a large magnet, or by the presence of scanner noise. Most likely, observed group difference was due to sampling error; never the less this demonstrates that, at minimum, testing inside the fMRI scanner environment did not interfere with task performance. Presented experimental designs could provide a future vehicle to investigate plasticity and training effects in ASA with highly trained musicians. In order to perform experiments in an equally challenging ASA environment as for non-musicians, task design could be adjusted to include melodies synthesized by the same instrument which incorporate incorrect and/or incomplete cues in only one stream for the aggregate condition, mis-tuned notes within triplets, or incomplete triplet patterns consisting of two or three triplet notes. Further differences may exist within musicians based on their specific training, with soloists possibly showing more difficulty with source segregation than conductors or orchestra members who are constantly separating their own instruments in the presence of multiple competing music streams, and a more general enhanced perceptual segregation of their main performing instrument (for example, Carey et al., 2015; Pantev et al., 2001). Further understanding of both subject's locus of attention and polyphonic listening behavior could be achieved by employing trials which contain a non-matching instruction and response, for example an instruction indicating attention to the aggregate and a response whether or not triplets were present in the bassoon.

Observed learning effects in subjects may be specific to the pitch ranges and instruments employed in the paradigm, not reflecting a general stream segregation performance gain which is transferable to other instruments or more common streaming tasks such as speech in noise detection. Within polyphonic music pieces, the upper voice is typically more salient (Crawley et al., 2002; Palmer & Holleran, 1994), suggesting that the segregation task may be more difficult when

attending to the cello. Even though no performance difference is observed between the bassoon and cello condition in either of the experiments, verbal reports do confirm that subjectively listeners found the cello condition more difficult. In an attempt to control for the influence of timbre on the learning effects, timbres of voices could be switched to provide an indication whether learning was specific for the instrument-pitch relationship or resulted in general music streaming improvements. To try and reduce some specificity effects of learning, both the training and experimental melodies were uniquely written with a maximum variation in melody and pitch to maintain recognition as common polyphonic music while not destroying the similar pitch relation necessary amid voices across compositions. It is currently unknown whether performance on a music stream segregation task reflects general stream segregation or is more specific to music. A better understanding of their link could be achieved by performance comparison towards standardized complex sounds in background noise tasks such as speech in noise (for examples see, Wilson, 2003; Killion et al., 2004; Kalikow et al., 1977; Nilsson et al., 1994) or, in addition, a music-specific stream segregation measure such as the music in noise test (Coffey et al., 2017b). This would allow insight into whether music streaming employs similar mechanisms as the more widely established segregation task of isolating speech from background sounds. Subject performance will probably be very comparable on the music and speech in noise tasks, although an increase in subjects' musical training may cause larger performance gains on the music in noise paradigm compared to the speech in noise task.

Polyphonic Music Perception

Several cognitive theories have been proposed to explain attention to polyphonic music, even though its neural processes are relatively unknown (for examples see, Janata et al., 2002; Ragert et al., 2014). Two of the main competing hypotheses are a divided attention (Gregory, 1990) and a figure-ground model (Sloboda & Edworthy, 2016). The divided attention model explains superior performance on polyphonic tasks, as compared to, for example, speech, by listeners' apparent capacity to divide their attentional resources over multiple melodic lines (Gregory, 1990). The figure-ground model, on the contrary, proposes that listeners attend only to a single melody while all others are assigned to the background (Sloboda & Edworthy, 2016), achieving multi-voiced perception by shifting their locus of attention between scene elements, therefore explaining perception via undivided attention. Contrary to what these models suggest, subjects are probably not simply dividing or alternating attention, they develop strategies to counteract divided attention issues by allowing for a true integration of melodies (Bigand et al., 2000). Even though listening strategies slightly differ between non-musicians, who appear to only integrate the melodic lines into streams, and musicians who are capable of constantly switching their attention between the integration and segregation of melodic lines, the general integrative

model does appear to hold for both groups (Bigand et al., 2000), suggesting that attention to music may indeed differ from general auditory attention processes. The underlying neural attentional mechanism per se does probably not differ, it is the horizontal and vertical relationship which exists between melodies in combination with music-specific schema development which allows to both integrate and segregate music voices, further explaining musicians' superior performance on these tasks. Schema-based processes are developed on the basis of acquired knowledge and provide an additional form of top-down information important for stream formation (Bey & McAdams, 2002; Bregman, 1990), operating either in an attentive or pre-attentive mode, depending on task demands. Schemas have been shown to modulate music segregation performance in auditory scenes where integration is strongly driven by primitive (*i.e.*, bottom-up) processes (Bey & McAdams, 2002; Bregman, 1990). When, for example, performing a segregation task with two interleaved melodies, it has been demonstrated that prior presentation of the to-be-attended sequence aids subsequent separation performance, while a frequency-transposition of the melody caused a reduction of these effects (Bey & McAdams, 2003). Within the current task we opted to not implement a modulation of top-down cues and focused on bottom-up attentive effects only, even though these cues are of great importance in ASA (Bregman, 1990; Ciocca, 2008; McAdams & Bregman, 1979; Micheyl et al., 2007) and could be employed to both investigate their contribution to music streaming and further aid or impede both the segregation and integration performance. Taken together, the interaction of both bottom-up and top-down processes appears to be capable of modulating whether subjects perceive multi-voiced music as integrated or segregated.

These considerations regarding musical stream segregation and the possible distinct mechanisms that are at play during music listening are also relevant for a broader understanding of how musical training may influence auditory cognition. For example, considerable evidence indicates that musicians outperform those without training in speech-in-noise tasks (Parbery-Clark et al., 2009; Zendel et al., 2015; Swaminathan et al., 2015, for review, see Coffey et al., 2017c). The neural mechanisms underlying this enhancement are not fully understood, even though there is evidence that both bottom-up mechanisms, centered within brainstem nuclei and auditory cortices (Bidelman et al., 2014; Coffey et al., 2017a), and top-down mechanisms (Kraus & Chandrasekaran, 2010), engaging motor and frontal-lobe systems (Du & Zatorre, 2017), play a role due to music's reliance on both kinds of processes. The task presented here could be used in conjunction with other tasks requiring segregation of targets from backgrounds to generate a better understanding of the relationship between music-specific auditory cognitive abilities, and their possible generalization to non-musical contexts.

Conclusion

In this work we demonstrated that participants with limited to no musical education could be trained to isolate triplet patterns in polyphonic music, and complete both a selective attention task and a combined attention-timbre manipulation auditory streaming task with high accuracy. Triplet detection provided us with an objective variable assessing the listener's locus of attention, as well as their general task compliance, showing they were able to both successfully segregate individual instruments and integrate across the music streams. Insight into ASA processes with long complex music stimuli could be employed to inform research into, among others, hearing aid design and brain-based algorithm development for hearing aids, Brain Computer Interfaces, and provide a powerful means to investigate the neural mechanisms underlying both stream segregation and integration in naturalistic though well-controlled auditory scenes. An understanding of general ASA processes in the brain may very well be one of the necessary hurdles to cross in order to discern those processes underlying general music processing (Nelken, 2008).

References

Alain, C. & Bernstein, L. J. (2015). Auditory Scene Analysis: Tales from Cognitive Neurosciences. *Music Perception: An Interdisciplinary Journal*, 33(1), 70–82.

Amaro, E., Williams, S. C. R., Shergill, S. S., Fu, C. H. Y., MacSweeney, M., Picchioni, M. M., Brammer, M. J., & McGuire, P. K. (2002). Acoustic noise and functional magnetic resonance imaging: Current strategies and future prospects. *Journal of Magnetic Resonance Imaging*, 16(5), 497–510.

Andoh, J., Ferreira, M., Leppert, I. R., Matsushita, R., Pike, B., & Zatorre, R. J. (2017). How restful is it with all that noise? Comparison of Interleaved silent steady state (ISSS) and conventional imaging in resting-state fMRI. *Neuroimage*, 147(C), 726–735.

Belin, P., Zatorre, R. J., Hoge, R., Evans, A. C., & Pike, B. (1999). Event-related fMRI of the auditory cortex. *Neuroimage*, 10(4), 417–429.

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Emerson, R. G., & Schroeder, C. E. (2011). Tuning of the Human Neocortex to the Temporal Dynamics of Attended Events. *The Journal of Neuroscience*, 31(9), 3176–3185.

Bey, C. & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics*, 64(5), 844–854.

Bey, C. & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 267–279.

Bidelman, G. M., Weiss, M. W., Moreno, S., & Alain, C. (2014). Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *European Journal of Neuroscience*, 40(4), 2662–2673.

Bigand, E., Foret, S., & McAdams, S. (2000). Divided attention in music. *International Journal of Psychology*, 35(6), 270–278.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The Perceptual Organization of Sound. MIT Press.

Bregman, A. S. (2015). Progress in Understanding Auditory Scene Analysis. *Music Perception: An Interdisciplinary Journal*, 33(1), 12–19.

Bregman, A. S. & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1), 19–31.

Brochard, R., Drake, C., Botte, M. C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1742–1759.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, 137(C), 81–105.

Carlyon, R. P. (2003). How the brain separates sounds. *Trends in Cognitive Sciences*, 8(10), 465–471.

Carlyon, R. P. & Cusack, R. (2005). Effects of Attention on Auditory Perceptual Organization. In L. Itti, G. Rees, & J. K. Tsotos (Eds.), *Neurobiology of Attention* (pp. 317–323). Cambrige, MA: Elsevier.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience-Landmark*, 13(13), 148–169.

Coffey, E. B. J., Chepesiuk, A. M. P., Herholz, S. C., Baillet, S., & Zatorre, R. J. (2017a). Neural Correlates of Early Sound Encoding and their Relationship to Speech-in-Noise Perception. *Frontiers in Neuroscience*, 11, 177–14.

Coffey, E. B. J., Lim, A. R., & Zatorre, R. J. (2017b). The Music-In-Noise Task: a tool for dissecting complex auditory perception. In *The Neurosciences and Music VI Music, Sound, and Health* Boston.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017c). Speech-in-noise perception in musicians: A review. *Hearing Research*, 352, 49–69.

Coffey, E. B. J., Scala, S., & Zatorre, R. J. (2011). Montreal Music History Questionnaire: a tool for the assessment of music-related experience. In *Neurosciences and Music IV Learning and Memory* Edinburgh, UK.

Crawley, E. J., Acker-Mills, B. E., Pastore, R. E., & Weil, S. (2002). Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), 367–378.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656.

Cusack, R. & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, 62(5), 1112–1120.

Deutsch, D. (2010). Hearing music in ensembles. *Physics Today*, 63(2), 40–45.

Deutsch, D. (2013). Grouping Mechanisms in Music. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 183–248). London, UK: Elsevier.

Dietz, L. (2010). Directed factor graph notation for generative models. *Max Planck Institute for Informatics, Tech. Rep.*

Du, Y. & Zatorre, R. J. (2017). Musical training sharpens and bonds ears and tongue to hear speech better. *Proceedings of the National Academy of Sciences*, 114(51), 13579–13584.

Gregory, A. H. (1990). Listening to Polyphonic Music. *Psychology of Music*, 18(2), 163–170.

Gregory, A. H. (1994). Timbre and auditory streaming. *Music Perception: An Interdisciplinary Journal*, 12(2), 161–174.

Hall, A. J., Brown, T. A., Grahn, J. A., Gati, J. S., Nixon, P. L., Hughes, S. M., Menon, R. S., & Lomber, S. G. (2014). There's more than one way to scan a cat: Imaging cat auditory cortex with high-field fMRI using continuous or sparse sampling. *Journal of Neuroscience Methods*, 224, 96–106.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective, & Behavioral Neuroscience*, 2(2), 121–140.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the acoustical society of America*, 61(5), 1337–1351.

Kawahara, H. & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *International Conference on Acoustics, Speech, and Signal Processing, 2003.* (pp. 256–259).: IEEE.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 116(4 I), 2395–2405.

Kraus, N. & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, (pp. 1–7).

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis*. A Tutorial with R, JAGS, and Stan. London, UK: Academic Press, 2 edition.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, 77(4), 750–761.

Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, 30(3), 259–274.

McAdams, S. (2013a). Musical timbre perception. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 35–68). London, UK: Elsevier Inc.

McAdams, S. (2013b). Timbre as a structuring force in music. In *ICA 2013 Montreal* (pp. 1–6).: ASA.

McAdams, S. & Bregman, A. S. (1979). Hearing Musical Streams. *Computer Music Journal*, 3(4), 26–43.

McDermott, J. H. & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology*, 18(4), 452–463.

Melara, R. D. & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics*, 48(2), 169–178.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., & Courtenay Wilson, E. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research*, 229(1-2), 116–131.

Moore, B. C. J. & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica united with Acustica*, 88(3), 320–333.

Nelken, I. (2008). Neurons and objects: the case of auditory cortex. *Frontiers in Neuroscience*, 2(1), 107.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the acoustical society of America*, 95(2), 1085–1099.

Palmer, C. & Holleran, S. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception & Psychophysics*, 56(3), 301–312.

Pantev, C., Roberts, L. E., Schulz, M., & Engelien, A. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport*, 12(1), 169–174.

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician Enhancement for Speech-In-Noise. *Ear and Hearing*, 30(6), 653–661.

Peretz, I. & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology*, 56(1), 89–114.

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and Integration of Auditory Streams when Listening to Multi-Part Music. *PLoS ONE*, 9(1), 1–9.

Riecke, L., Peters, J. C., Valente, G., Kemper, V. G., Formisano, E., & Sorger, B. (2016). Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cerebral Cortex*, advance online access, 1–13.

Shamma, S. A. & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3), 361–366.

Sloboda, J. & Edworthy, J. (2016). Attending To Two Melodies At Once: the of Key Relatedness. *Psychology of Music*, 9(1), 39–43.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137–149.

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the acoustical society of America*, 117(3), 1285–14.

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, 69(1), 136–152.

Swaminathan, J., Mason, C. R., Streeter, T. M., Best, V., Gerald Kidd, J., & Patel, A. D. (2015). Musical training, individual differences and the cocktail party problem. *Scientific reports*, (pp. 1–11).

Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *Neuroimage*, 77, 52–61.

van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America*, 61(4), 1041–1045.

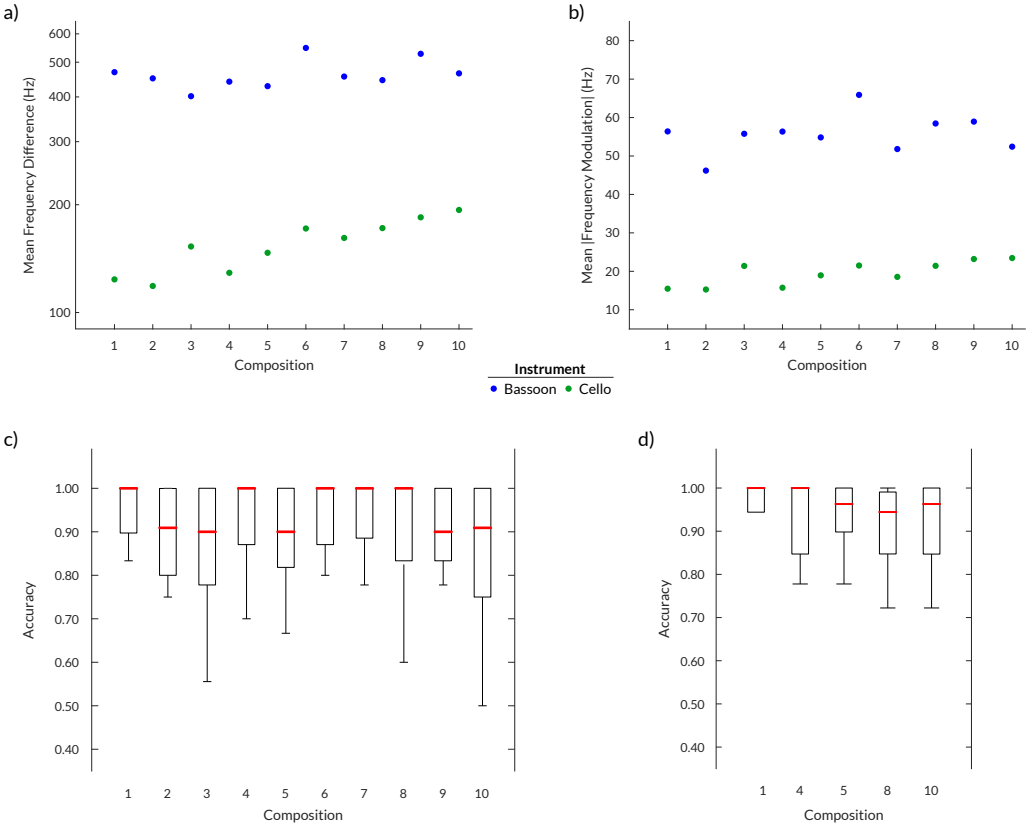
Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2), 45–52.

Wilson, R. H. (2003). Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *Journal of the American Academy of Audiology*, 14(9), 453–470.

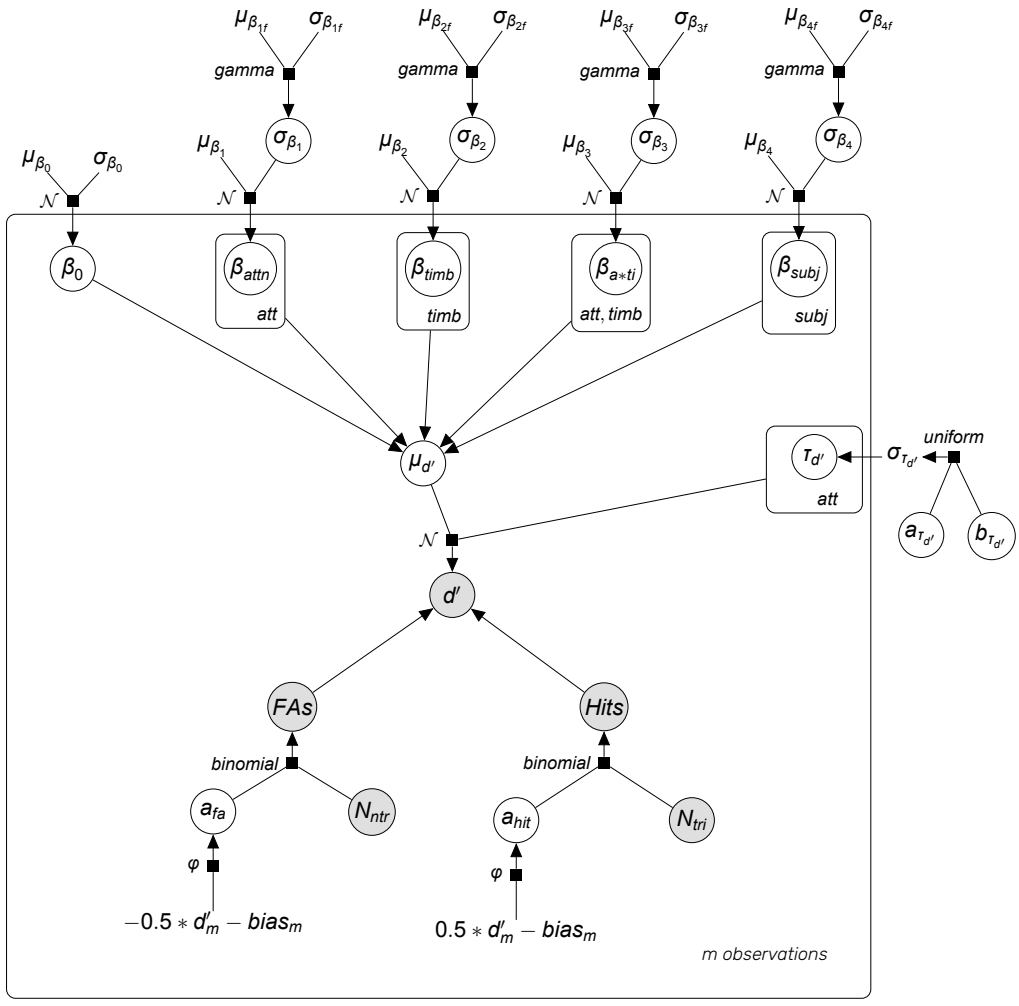
Zatorre, R. J. & Zarate, J. M. (2012). Cortical Processing of Music. In *The Human Auditory Cortex* (pp. 261–294). New York, NY: Springer New York.

Zendel, B. R., Tremblay, C.-D., Belleville, S., & Peretz, I. (2015). The Impact of Musicianship on the Cortical Mechanisms Related to Separating Speech from Background Noise. *Journal of Cognitive Neuroscience*, 27(5), 1044–1059.

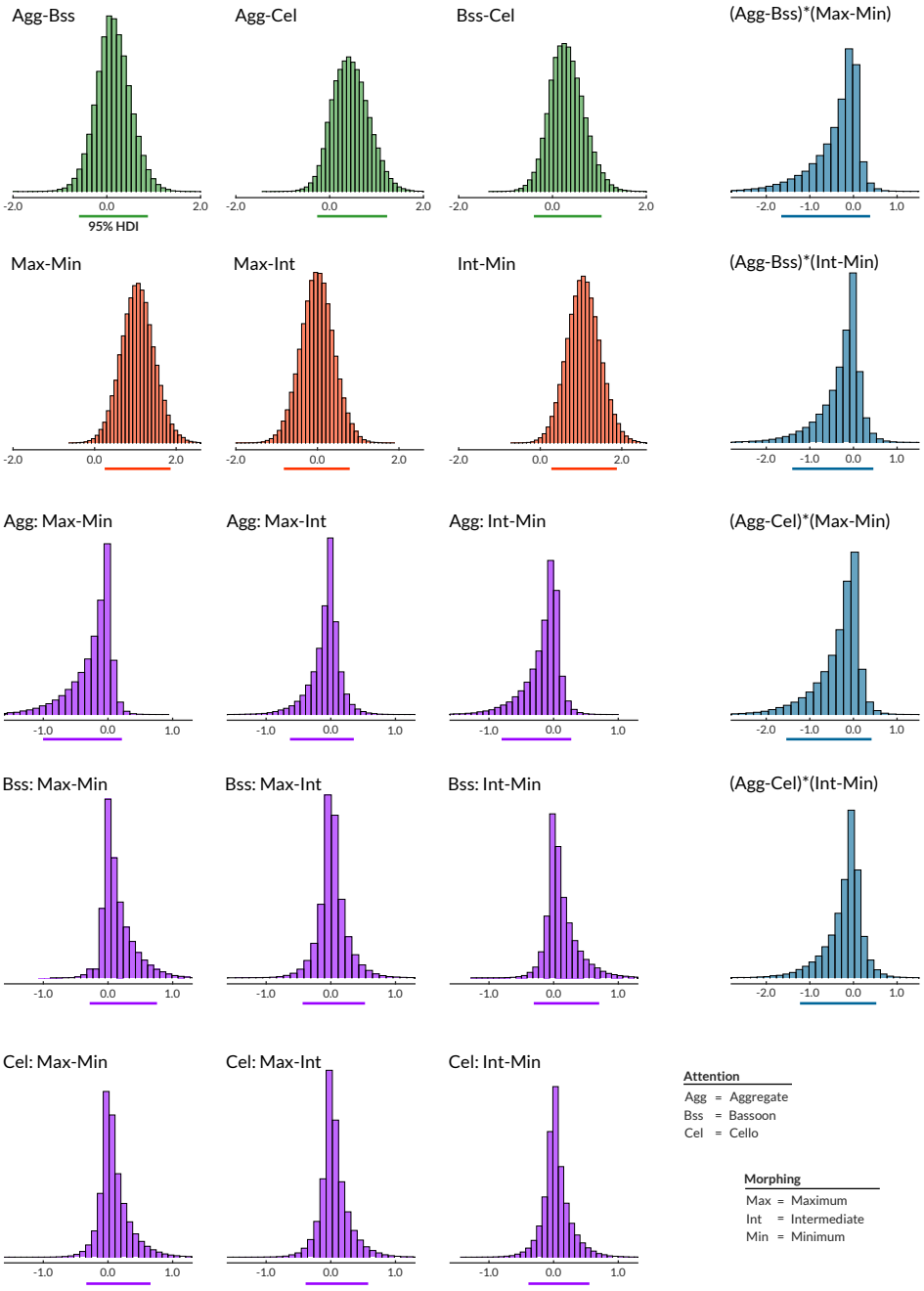
Supplementary Figures



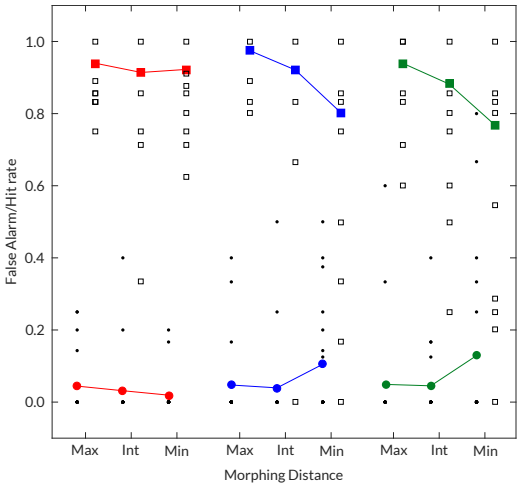
Supplementary Figure 2.1: Composition Details and Accuracies. Mean pitch (a) and mean absolute frequency modulation (b) for bassoon (blue dots) and cello (green dots) per composition included in Experiment 1. Accuracy per composition across all trials for experiment 1 (c) and experiment 2 (d).



Supplementary Figure 2.2: Experiment 2 Directed Acyclic Graph. Bayesian hierarchical model parameters for the d-prime model. Bias is modeled equally to d-prime, connecting to this model within the Hit and FA nodes. For simplicity, bias modeling is omitted from this graph. Notation based on Dietz (2010); see http://github.com/jluttine/tikz-bayesnet



Supplementary Figure 2.3: Experiment 2 Bayesian Hierarchical Model Effects. Several possible contrasts and interactions; horizontal lines below histograms, 95% HDI



Supplementary Figure 2.4: Experiment 2 Condition Specific Signal Detection Results. Hit rates (squares) and False Alarm rates (dots) for group mean (colors) and individuals (black). Attention condition: red = aggregate, blue = bassoon, and green = cello; Timbre distance: max = maximum, int = intermediate, min = minimum.

"As long as our brain is a mystery, the universe, the reflection of the structure of the brain will also be a mystery"

Santiago Ramón y Cajal

3

Segregation or integration of polyphonic music modulates cortical auditory response patterns

Based on: Disbergen, N. R., Valente, G., Zatorre, R. J., and Formisano, E. (to be submitted). Segregation or integration of polyphonic music modulates cortical auditory responses patterns.

Abstract

Music is typically perceived by integrating the various instruments present in the piece. However, under attentional control it is alternatively possible to selectively attend to individual instruments. We investigated the neural correlates of this top-down modulation for auditory stream segregation and integration with functional MRI at 7 Tesla, using a previously validated behavioral paradigm. Nine non-musicians listened to custom-composed polyphonic music comprising a bassoon and cello timbre under two conditions: attending to the individual instruments or to the aggregate, all while detecting triplet patterns in the music as a control for task performance. Data were analyzed via a novel combination of independent component analysis and multivoxel pattern analysis techniques. Results indicated that the listener's attentional state (*i.e.*, integration or segregation) could be decoded above chance at the individual subject level within the frontal-temporal attention network, despite that the stimuli and behavioral responses were identical across tasks. Subsequent region of interest analysis demonstrated significant above-chance classification as early as in primary auditory areas. Findings support the hypothesis of early auditory area involvement in auditory feature integration and stream formation.

Introduction

When listening to a song of our favorite rock band or a classical music concert, we are typically unaware that what we are appreciating is the result of a carefully orchestrated combination from many different sound sources, *id est* the various musical instruments or lines of music. At times we do attempt to single out a specific instrument to enjoy, for example, a stunning solo of the lead guitarist or the virtuosity of the piano soloist. In this context, music listening provides an exemplary illustration of how our ongoing listening intention can modify the auditory processing and perception of complex mixtures of sounds (or auditory scenes), leading to either the perceptual integration or segregation of simultaneously present sound streams (Pressnitzer et al., 2011).

Psychophysical (e.g., Bregman, 1990; Brochard et al., 1999; Carlyon & Cusack, 2005; Cusack et al., 2004) and neuroscience (e.g., Besle et al., 2011; Carlyon, 2003; Elhilali et al., 2009; Lakatos et al., 2013; Shamma & Micheyl, 2010; Riecke et al., 2016; Sussman et al., 2007) research has shown that the extent to which simultaneous sound streams are segregated or integrated depends mostly on their physical (*i.e.*, acoustic) characteristics, such as their inter-relationship and degree of overlap in both time and frequency. This type of bottom-up processing of acoustic features interacts with and can be modulated by intention or attention driven top-down processes. For polyphonic (*i.e.*, multi-instrument) music, pitch and timbre differences between the instruments (*i.e.*, voices) provide prominent bottom-up cues for their segregation (e.g., Bregman & Pinker, 1978; Cusack & Roberts, 2000; Deutsch, 2013; Marozeau et al., 2013; McAdams, 2013b,a; Wessel, 1979). The structural relationship existing between the individual voices in polyphonic music, instead, promotes their integration into a single holistic percept of multiple simultaneously playing instruments (Bigand et al., 2000; Bregman, 1990; Gregory, 1990). Based on these characteristics, polyphonic music is well suited for the study of Auditory Scene Analysis (ASA) in complex and naturalistic scenarios, as typically encountered by listeners on a daily basis. Within a contrapuntal polyphonic music piece, the composer achieves instrument integration by employing specific melodic links between the individual notes both within a voice (*i.e.*, horizontal coherence) and between the voices (*i.e.*, vertical coherence). Contrary to the more traditional speech-based cocktail-party experiments which permit only the investigation of source segregation, polyphonic music additionally allows for the investigation of stream integration processes across sounds, which has been investigated only to a limited extent (Deutsch, 2013; Disbergen et al., 2018; Ragert et al., 2014; Sussman, 2005; Uhlig et al., 2013).

In the present study we exploit this unique quality of music and investigate the neural mechanisms underlying the segregation and integration of auditory streams within complex auditory

scenes using polyphonic music. While the neuroscientific investigation of music listening and processing per se has increased in recent years (for reviews see, McDermott & Oxenham, 2008; Peretz & Zatorre, 2005; Zatorre & Zarate, 2012), only a small number of studies have employed functional magnetic resonance imaging (fMRI) in combination with complex music stimuli to examine its attentional aspects. Ragert et al. (2014) demonstrated that in musicians, integration and segregation elicited differential cortical activations in the intraparietal sulcus (IPS) and the planum temporale (PT). Janata et al. (2002), showed that areas in the left fusiform gyrus and a multitude of occipital areas were modulated during segregation, whereas during integration changes were seen in bilateral parietal, right superior frontal, and left anterior cingulate cortex. In addition, they showed general task activation changes in the superior temporal gyrus (STG), intraparietal sulcus (IPS), and several frontal areas, similar to those involved in speech scene analysis. Evidence from these studies showed little convergence regarding the brain networks involved. Studies investigating the attentional modulation of complex music listening or scene analysis were mostly conducted in expert musicians, which have been shown to have modified listening behavior compared to non-musicians (e.g., Puschmann et al., 2018; Coffey et al., 2017).

Furthermore, the lack of effects at the level of early auditory cortex in previous experiments employing music, is in contrast with the majority of studies which have investigated the neural basis of auditory scene analysis by virtue of relatively schematic scene elements, such as multiple alternating tone sequences or tones in noise. Electrophysiology in animals has shown that already at the level of primary auditory cortex, top-down modulations of schematic stimuli can be observed (e.g., Atiani et al., 2009; Fritz et al., 2003), typically leading to an enhanced processing of the attended sound's feature(s). In humans, several fMRI studies suggested comparable effects in primary auditory cortex for simple auditory scenes containing tone-sequences (Da Costa et al., 2013; Paltoglou et al., 2009; Riecke et al., 2016).

Under more naturalistic listening conditions, ASA has often been investigated in the context of multi-talker environments (Snyder & Alain, 2007), combined with various neuroimaging methods these provided evidence for selective enhancement of the attended speaker's representation in early auditory cortex (e.g., Mesgarani & Chang, 2012; Golumbic et al., 2013; Ding & Simon, 2012; O'Sullivan et al., 2015; Puschmann et al., 2018). Modulations observed in early auditory areas probably originate in frontal areas, especially the ventral prefrontal cortex (Cohen et al., 2009; Hill & Miller, 2010; Ahissar et al., 2009; Atiani et al., 2014; Bizley & Cohen, 2013; Griffiths & Warren, 2004; Shinn-Cunningham, 2008; Hausfeld et al., 2018). Either with tone sequences or in multi-talker environments, most of the previous work has focused on neural mechanisms underlying the attentive selection (*i.e.*, segregation) of an individual stream within the auditory scene, while much less is known about those mechanisms facilitating the integration of auditory streams.

In the present study, we investigated the neural mechanisms of both stream integration and segregation by combining high-spatial resolution fMRI at 7 Tesla with a previously validated psychophysical music ASA paradigm employing custom-composed polyphonic pieces (Disbergen et al., 2018, Chapter 2). During fMRI measurements, listeners (N=9, non-musicians) were presented with two-voice polyphonic music pieces synthesized in bassoon and cello timbre. They were asked to detect a pattern of rhythmic modulations incorporated either within or across the two musical instruments. Patterns were located within the second half of the stimulus to stimulate attentional deployment during the majority of the segment (Fig. 3.1a). The locus of attention was changed between individual instruments (*i.e.*, segregation) and the aggregate (*i.e.*, both instruments; integration) by a visual instruction.

We expected that listening to the musical pieces while segregating or integrating the component voices would modulate activity in a frontal-temporal network of cortical areas typically involved in attentive (music) listening. In addition, we hypothesized that, in agreement with ASA studies employing tone sequences, these modulatory effects could be observed in early auditory cortex (*i.e.*, Heschl's gyrus). To allow for the discovery of such potentially subtle modulations, which typically remain undetected with massive univariate approaches, we investigated these hypotheses via a novel methodological approach which combined independent component analysis (ICA) for the unbiased definition of region of interests with multivariate decoding for the discrimination of activity patterns. In addition to this generalized network approach, we investigated decoding performance within a multitude of anatomical sub-regions of this frontal-temporal network, including Heschl's Gyrus.

Methods

Subjects

Nine right-handed adult volunteers (four women; age 23.6 +/- 2.4 years, mean +/- standard deviation) with self-reported normal hearing, motor, and vision abilities participated in this study. None of the subjects spoke a tonal language and all had less than two years of (formal) musical training on a lifetime basis, as assessed with the Montreal Music History Questionnaire (Coffey et al., 2011). Volunteers were mostly students recruited from Maastricht University and provided written informed consent in accordance with the protocol as approved by the Maastricht University Ethics Review Committee Psychology and Neuroscience.

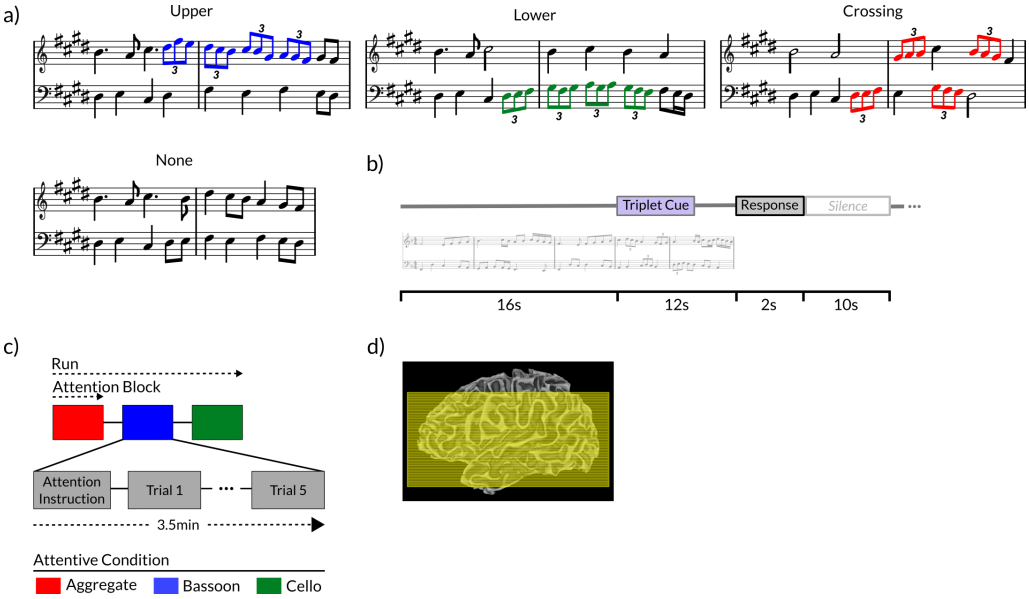


Figure 3.1: Design & Coverage. Different triplet versions for each music composition (a): no triplets, upper voice (*i.e.*, bassoon; blue notes), lower voice (*i.e.*, cello; green notes), crossing voices (red notes). Trial buildup (b) with 28s stimulus, 2s response window, and a 10s silence. Trials (c) were presented in attentive blocks of five trials each, preceded by a visual attention instruction and silence. Typical slice coverage (d) for functional MRI acquisitions.

Stimuli

In this experiment, we made use of five custom-composed two-instrument polyphonic counterpoint music pieces of 28 seconds (s) duration, synthesized in a bassoon (treble clef) or cello (bass clef) timbre at a tempo of 60 beats per minute*. Frequency ranges of the voices varied between compositions within a range of 82-277Hz for the lower and 220-880Hz for the upper voice; see Disbergen et al. (2018, Chapter 2) for further stimulus design details. Music was synthesized in mono for bassoon and cello independently, employing Musical Instrument Digital Interface (MIDI) files with Logic Pro 9 (Apple Inc., Cupertino, California, USA) and sampled at 44.1 kHz with a 16Bits resolution. Single-instrument files were Root Mean Square (RMS) equalized, combined into polyphonic pieces, their onsets and offsets ramped exponentially with a rise-fall time of 100ms, and filtered per individual channel with Sensimetrics (Sensimetrics Corporation, Malden, Massachusetts, USA) equalization filters in MATLAB (The MathWorks Inc., Natick, Massachusetts, USA). All stimulus processing and manipulation aside from synthesizing was performed via custom-developed MATLAB codes.

Listeners detected rhythmic modulations in the music, comprising a pattern of triplets which

*Experimental stimuli are available for download via the Zatorre lab's website: http://www.zlab.mcgill.ca

were carefully incorporated into the melodic structure as not to stand out from the surrounding music (Fig. 3.1a). Triplets consisted of three eighth notes each and were played in the time of one beat, which in non-musically trained individuals typically leads to the perception of a slight 'speeding-up' compared to the neighboring notes. This specific temporal modulation was selected due to its orthogonality with respect to pitch-based segregation mechanisms, further aiding its detection by non-musically trained listeners, as previously validated in Disbergen et al. (2018, Chapter 2). The patterns detected by subjects comprised four eighth-note triplets in a row, with a total duration of four seconds. Patterns followed the excerpt's melody and could be located within the bassoon (Fig. 3.1a, blue notes), the cello (Fig. 3.1a, green notes), across voices (Fig. 3.1a, red notes), or not present. When the triplets crossed voices, they initiated randomly with the first triplet in the bassoon or cello, and accordingly alternated between the voices. When patterns were located within a single music voice, they were only present in the respective instrument's voice. Starting time of the triplet patterns was pseudo-randomly distributed in the second half of the stimulus between 16 and 19 seconds, resulting in a stimulus set which was physically identical up to 16 seconds and only differed as to the inclusion and position of triplets.

Design and Behavioral Analysis

In order to gain insight into the subject's locus of attention during scanning, participants performed a forced-choice delayed-response target detection task within a single musical voice (bassoon or cello) or integrated across voices. Stimuli were presented in blocks of five trials each under the same attentive condition (Fig. 3.1c). A trial contained the stimulus, a response window of 2s, and a 10s post-stimulus silence. Before the beginning of each attention-block, listeners were visually instructed to attend to the bassoon, the cello, or the aggregate (*i.e.*, both instruments), which was followed by a 15s silence. On each trial subjects responded via a post-stimulus button press whether or not the pattern of triplets was present in the instrument(s) they had been instructed to attend to. Within a run 15 trials were presented, covering all three attention conditions, and containing eight target and seven control trials. A total of 135 trials was presented across nine functional runs, hence the full stimulus set was presented three times, albeit with a unique stimulus order for each repetition.

Triplet presence and/or location differed depending on the condition: when attending both instruments, target trials contained a pattern crossing voices and control trials had no triplets; in the attend bassoon condition, target trials contained a pattern in the bassoon, control trials contained a pattern in the cello or no triplets; similarly, when attending cello the target trials contained a pattern in the cello and control trials included a pattern in bassoon or no triplets. Due to the employment of seven control trials per run, bassoon and cello conditions contained an uneven number of trials with triplets in the unattended instrument or with no triplet present.

Within the current experiment, an additional modulation of instrument timbre across three discrete values was included (see Disbergen et al., 2018, Chapter 2, Experiment 2 for details): each melody was presented in its original timbre (*i.e.*, no manipulation), a perceptually determined minimum timbre difference between instruments, and a third intermediate distance which was 20% closer to the maximum distance measured from the minimum distance. For improved sensitivity, we have opted to analyze the attentive modulation by inclusion of all timbre distance trials.

Stimulus presentation order was pseudo-random, controlling that within an attention block of five trials, compositions could only occur once. Within a run, stimuli could only occur once, the first stimulus of a consecutive attentive block could not be the same as the last of the previous, and each condition could occur only once. Each timbre distance version of each composition was covered by all conditions over the course of three runs/nine attentive blocks, while within an attention block the same timbre distances could not follow, a composition's timbre distance had to occur at least once, and the same timbre distance could not occur more than twice. Condition order across three runs (*i.e.*, one experiment repetition) had to be unique, with each condition appearing once at each position within the three-block sequence, for example: ABC-BCA-CAB. Within a participant across all three experiment repetitions (*i.e.*, nine runs) the condition order blocks appeared in all positions, for example: repetition 1) ABC-BCA-CAB, 2) BCA-CAB-ABC, and 3) CAB-ABC-BCA. Across subjects the condition run order was further balanced, for example: participant 1) BCA-CAB-ABC versus 2) CBA-BAC-ACB. Stimulus presentation and response recording was carried out with Presentation 17.0 (Neurobehavioral Systems Inc., Albany, California, USA), employing Sensimetrics S14 ear-buds with foam-tips providing hearing protection, and a Creative Sound Blaster X-Fi Xtreme Audio (Creative Technology Ltd., Singapore) sound card at approximately 94 dB SPL.

To allow non musically-trained individuals to participate in the experiment and achieve adequate task performance, they were subjected to a training session. Training initiated with scales including triplets and concluded with complex experiment-level melodies which had the triplet pattern incorporated. During each step, instruction and examples were provided, followed by several test-stimuli to evaluate learning. Examples could be repeated as often as desired and training protocols were automatically adapted to participant performance. At the end of a training session, generalization was tested via a pre-test which simulated an abbreviated version of the experiment, including 24 trials across 4 unique compositions, scoring a minimum of 85 percent correct; all subjects completed training successfully. Any music employed outside of the main experiment was custom written for these purposes and unrelated to the experimental melodies. For an in-depth discussion of the task and training, along with a demonstration of its validity, including its use with non-musically trained subjects inside the MRI scanner environment, please

refer to Disbergen et al. (2018, Chapter 2).

Subject behavioral responses were classified as hits H, misses M, false alarms FA, and correct rejections CR. Due to scores occurring close to or at ceiling, their d-prime values were edge-corrected (e.g., Stanislaw & Todorov, 1999):

$$d'_{ec} = \Phi^{-1}(\frac{H + 0.5}{H + M + 1}) - \Phi^{-1}(\frac{FA + 0.5}{FA + CR + 1}) \tag{3.1}$$

where Φ^{-1} denotes the inverse normal cumulative distribution function with $\mu = 0$ and $\sigma = 1$.

MRI Acquisition

Anatomical and functional data were acquired on a 7 Tesla Siemens Magnetom MRI system employing a Nova Medical head radiofrequency coil (single transmit, 32 receive channels; Nova Medical Inc., Wilmington, MA, USA). Anatomical T1-weighted images were collected at a 0.65 mm isotropic resolution with a modified Magnetization Prepared 2 Rapid Acquisition Gradient Echoes (MP2RAGE; Marques et al., 2010) sequence (240 sagittal slices, repetition time (TR) = 5000ms, inversion time 1 (TI1)/TI2 = 900/2750ms, flip angle 1 (FA1)/FA2 = 5/3 degrees, echo time (TE) = 2.51ms, generalized autocalibrating partial parallel acquisition (GRAPPA) = 3, acquisition time (TA) = 10:17 minutes). Functional T2*-weighted data were measured at a 1.2mm isotropic resolution using a continuous multi-band gradient echo echo planar imaging sequence (mb-EPI; TR = 2000ms, TE = 19ms, GRAPPA = 3, Multi-band = 2, partial Fourier = 6/8, anterior to posterior phase encoding, 60 interleaved transverse slices per volume with no inter-slice gap, 354 volumes per run, TA = 12:10 min; Feinberg et al., 2010; Moeller et al., 2010; Setsompop et al., 2012). A functional volume covered the full brain transversally, while limiting coronal coverage by excluding part of the dorsal-parietal lobe and a section of anterior-dorsal frontal lobe (Fig. 3.1d). In order to allow post-hoc estimation and correction of susceptibility-induced off-resonance field distortions, five reverse phase-encoded volumes were measured at the beginning of each functional block within a session as well as after each individual run. Functional acquisitions were divided over a total of nine runs, acquired across two (7 subjects) or three (2 subjects) scanning sessions of about three hours each on separate days.

Anatomical Analysis

T1-weighted (UNI) MP2RAGE images were employed for both functional alignment and segmentation purposes; all processing was performed with BrainVoyager 20.2 (Brain Innovations, Maastricht, The Netherlands) and custom MATLAB codes. UNI images were first masked with the second inversion time (TI2) image to remove noise outside of the tissue, and transformed (sinc

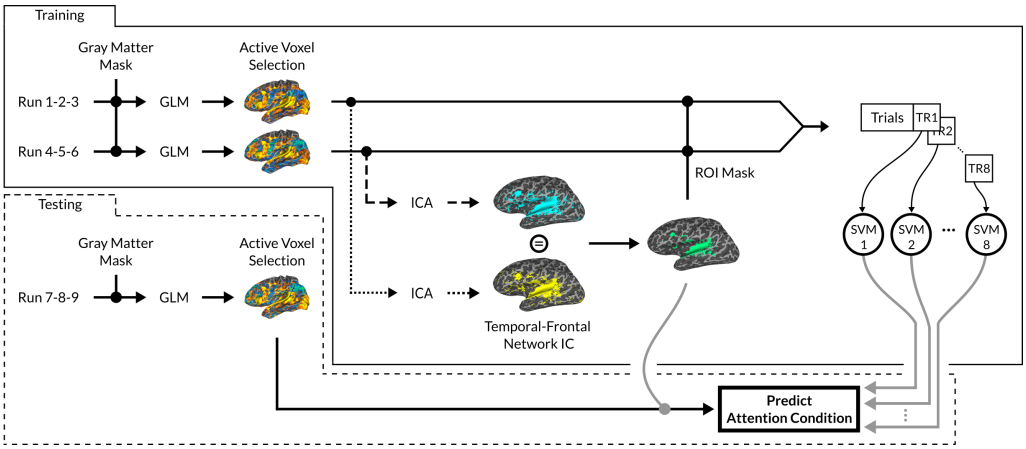


Figure 3.2: Functional Analysis Overview. Data were split into three sets of three runs (*i.e.*, one experiment repetition), and a GLM was fitted for each split withing a GM mask. A split's active voxel's Betas (FDR, q=.05) were injected into an ICA analysis, and for each split the Temporal-Frontal Network component was extracted. IC-based ROIs were generated by selecting only those IC voxels intersecting across the two training data splits. SVM models were fitted per TR for the first eight triplet-free time points.

interpolation) from 0.65mm isotropic to a 0.60mm isotropic resolution, corresponding to half the functional resolution. For each subject, a reference UNI image was selected from one of their sessions and ACPC-transformed (sinc interpolation), to which the UNI images of the other sessions were aligned via multi-scale intensity-driven alignment. When image quality allowed, the UNI images from multiple sessions were averaged in ACPC space to further improve data quality (7 subjects). Following the skull-stripping procedure, tissue contrast was enhanced by applying Gaussian smoothing with image boundary preservation. Due to data quality issues, subject S9's UNI reference image was further noise-reduced with FSL's SUSAN non-linear noise filter (Smith & Brady, 1997).

Cerebellums were manually removed from the volume, followed by gray matter (GM) to white matter (WM) boundary estimation via adaptive region growing. To reduce noise around the WM-GM boundary it was polished, after which sub-cortical structures and ventricles were manually tagged as WM and WM-GM boundaries adjusted where necessary. The GM was estimated by dilating WM borders until the cerebral spinal fluid (CSF), followed by a polishing of the GM-CSF border. GM masks were generated, to later be applied to the functional data, via GM thickness measures employing the Laplace method (Jones et al., 2000) as implemented in BrainVoyager. Anatomical regions of the Temporal lobe were defined on the GM-WM boundary inflated cortical mesh, employing individual anatomical markers in the volume based on the parcellation scheme by Kim et al. (2000) and considerations discussed in Moerel et al. (2014). For each individual brain and hemisphere, six anatomical ROIs (Fig. 3.6a) were defined: Heschl's Gyrus (HG), Planum Polare (PP), Planum Temporale (PT), anterior Superior Temporal Gyrus (aSTG), medial

STG (mSTG), and posterior STG (pSTG). Volumetric representations of these areas were generated by a 5mm expansion along the vertex normals, followed by a masking of these projections to the GM volume only. Possible overlap between regions in the volume was resolved by assigning overlapping-voxels to a single region based on a ranking system, selecting the highest rank of the pair in the following order: HG, PP, PT, mSTG, pSTG, aSTG.

Functional Analysis

Overview

In order to identify networks of voxels responding similarly to attentive music listening, we divided the nine functional runs into their three respective experiment repetitions: 1-2-3, 4-5-6, and 7-8-9 (Fig. 3.2). Data were separated into a training and test set based on the experiment repetitions, for example training with runs 1-2-3 and 4-5-6, while testing with runs 7-8-9. A separate GLM was fitted for each of the three experiment repetitions within the individually defined GM-masks, selecting only active voxels for further analysis based on a false discovery rate (FDR) with q=.05. Remaining voxels time series were employed to perform blind source separation on each experiment repetition separately by means of Independent Component Analysis (ICA; Formisano et al., 2004; McKeown, 2003). From the resulting ICs, for each training-data repetition only, the components representing subject's frontal-temporal networks were considered and their union was used as the functionally defined frontal-temporal mask of that given training-set. The test-dataset was entirely left out of this procedure, hence preventing bias in the decoding results.

Employing a within-subject cross-validated setup we investigated whether this individually-defined frontal-temporal network was modulated by task. Initially, a multivariate classifier was trained on a TR-basis for the integration versus segregation classes within the entire IC-defined network of the training-data. Generalization performance was tested by masking the testing data with the training-data defined frontal-temporal network and predicting the attentive conditions of the before unseen test-data. Accordingly, to gain further insight into those anatomical areas displaying different response patterns, the attention task was classified within individually defined temporal anatomical ROIs as well as a 'difference-ROI' which contained those voxels not included in any of the anatomical definitions. Classification significance was assessed via within-subject permutation testing and multiple comparison corrections across TRs based on their empirical null-distributions; population-level inference was implemented based on the prevalence metric (Allefeld et al., 2016).

GLM and Response Estimation

Functional MRI data were preprocessed with BrainVoyager QX 2.8 and custom MATLAB codes, consisting of slice scan time correction based on the slice time acquisition table (sinc interpolation) and three-dimensional motion correction towards the first functional volume of the session (trilinear detection and sinc interpolation). Reversed phase-encoded EPI volumes were employed to estimate and correct susceptibility-induced off-resonance field distortions with the Topup algorithm implemented in FSL (Andersson et al., 2003; Smith et al., 2004), followed by temporal high-pass filtering at 11 cycles per run. Images were accordingly co-registered to their session-specific UNI image in native space via gradient-based affine alignment, allowing translation, rotation, and scaling adjustments; results were inspected and manually adjusted when necessary. Resulting images were transformed from native into ACPC space (sinc interpolation), as defined by the reference-session anatomical image (see Anatomical Analysis), followed by manual alignment inspection and correction if necessary. Resulting functional images in ACPC space were spatially smoothed with a Gaussian of 2.0mm full width at half maximum, based on considerations discussed in Gardumi et al. (2016). The nine runs were split into their respective balanced experiment repetitions of three runs each (1-2-3, 4-5-6, and 7-8-9; Fig. 3.2). GLMs were fitted for each split of three runs separately, employing a FIR model including both instructions and trials, *id est* each data-point was fitted with a separate stick predictor, adding additional three predictors at the end of the stimulus to improve modeling. Each GLM was computed within the GM mask only, applying a percent-signal-change transformation on a per-run basis. Within each run-split, only active voxels (FDR-based, q=.05) were selected and an ICA (60 ICs) performed on their respective GLM-Betas matrix.

ICA-based Pattern Selection

All ICAs were computed on temporally z-scored data with the FastICA algorithm (version 2.5; Hyvarinen, 1999), employing a deflation decorrelation approach and hyperbolic tangent nonlinearity (a=1). From each split's ICA, the subject's temporal-frontal component was extracted. In order to perform pattern-classification based inference and avoid a circular procedure which may introduce bias, any further analysis was performed in cross-validation of run-splits. For each fold, two run-splits (e.g., 1-2-3 and 4-5-6) were selected for training purposes and the intersection of their temporal-frontal ICs set as the functionally-defined ROI for that specific fold; see Fig. 3.3 for an overview of these ROIs and their overlap per subject. The third run-split (e.g., 7-8-9) was separated for testing purposes, and masked with the fold-specific temporal-frontal ROI as defined by the training-set (Fig. 3.2). Anatomical ROIs for classification were defined by the intersection between the temporal-frontal ROI and the anatomically defined ROI (see Anatomical Analysis). All functional assessments were performed in the pre-triplet occurrence window only,

id est time-point 1-8, during which stimuli were physically identical independent of their triplet version (see Stimuli).

Multivariate Classification & Performance Evaluation

Binary classification was performed with integration (aggregate condition) versus segregation (mix of attention to bassoon and attention to cello conditions) trials. Adopting a distribution including all trials would lead to largely unbalanced classes, therefore aggregate condition trials were classified against a balanced random sample containing an equal number of bassoon and cello condition trials, which resulted in 60 training-trials and 30 test-trials per fold; number of trials for S4 varied slightly due to exclusions based on motion artifacts, depending on the split combinations number of training trials was 54 or 56 and testing trials 26 or 28. Random trial selection for the segregation class was repeated 10 times per fold and classification accuracies were averaged across the repetitions. The adopted analysis strategy (Fig. 3.2) resulted in three unique functionally defined ROIs, within which all possible combinations of train and test sets were classified; for example, ROI estimation and training on run-split 1 and 2 (*i.e.*, run 1-2-3 and 4-5-6), and testing on trials of run-split 3 (*i.e.*, run 7-8-9). Training sets always included the run-splits from which the respective ROI was estimated, maintaining full independence of the test set. Training trials were temporally z-scored across TRs, while testing trails were z-scored using the training trials mean and standard deviation. The analysis was conducted with a binary linear Support Vector Machine (SVM) model, fitted per TR for all training trials. Following training, model's generalization performance was tested on a per-TR basis employing the independent testing data. Testing performance is reported as percent accurate, computed by averaging classifier accuracy over all folds within a given TR.

True-label classification accuracy divergence from chance was assessed at the single subject level by permutation testing (Golland et al., 2005), randomly scrambling trial labels 1000 times for each random selection of integration and segregation trials. All permutations were performed with the same data and in an identical manner as for true-label classification; permutations were performed at the level of the run-presentation order. For each permutation n ($n = 1, 2, \ldots, N$) the accuracy observed under the null was computed at each TR m ($m = 1, 2, \ldots, M$). The p-value associated with observed accuracy a_m at a given subject and TR was then computed as:

$$p_m = \frac{\sum_{i=1}^{N} I\left(\tilde{a}_m^{(i)} \geq a_m\right) + 1}{N + 1} \tag{3.2}$$

where a_m is the observed classification accuracy at TR m, $\tilde{a}_m^{(i)}$ denotes the observed accuracy obtained at the i-th permutation at the same TR m and the indicator function $I(A)$ is 1 if event A

occurs, otherwise it is zero.

To correct for multiple comparisons (*i.e.*, multiple TRs) within each subject, we employed a maximum permutation statistic correction (Nichols & Holmes, 2002), with $\alpha = .05$. For each permutation we considered the maximum value observed across all TRs

$$\tilde{a}^{(i)} = \max\left(\tilde{a}_1^{(i)}, \tilde{a}_2^{(i)}, \ldots, \tilde{a}_M^{(i)}\right) \tag{3.3}$$

and computed the corrected p-value p_m^* of a TR m as:

$$p_m^* = \frac{\sum_{i=1}^{N} I\left(\tilde{a}^{(i)} \geq a_m\right) + 1}{N+1} \tag{3.4}$$

With this approach, we could therefore identify within a single subject those time-points where decoding accuracy significantly differed from chance.

For group-level inference we did not constrain subjects to the same temporal decoding profile, instead we constructed a surrogate measure for each subject which was accordingly employed for population inference, for which we used two different approaches. First, a partial conjunction approach computed via the method suggested in Heller et al. (2007), which combines the p-values obtained at each TR (see eq. (3.2)) to produce an aggregated p-value associated with the null hypothesis that none of the subject's time-points showed a significant classification accuracy. Since the Blood Oxygen Level Dependent (BOLD) responses at consecutive time-points are highly correlated, we cannot assume that the p-values at different TRs are independent, and therefore made use of a generalization of the Simes p-value correction (Heller et al., 2007) for each individual subject, performing a partial conjunction whether at least one out of eight time-points showed an effect. The correction proposed by Heller et al. (2007) results in:

$$p^{u/m} = \min\left\{(m-u+1)p_m^{(u)}, \frac{(m-u+1)}{2}p_m^{(u+1)}, \ldots, \frac{(m-u+1)}{m-u}p_m^{(m-1)}, p_m^{(m)}\right\} \tag{3.5}$$

where $p_m^{(u)}$ are the sorted p-values across the time points, with $p_m^{(1)}$ being the smallest. Setting $u = 1$ and $m = 8$, resulting p-value is associated with the partial conjunction null that no time point has an effect, and, therefore, the alternative that *at least* 1 time point out of 8 has a true effect.

Second, we considered another maximum statistics approach, testing the maximum performance of each subject across TRs versus the maximum of each permutation, which resulted in a temporally unspecific p-value against the null hypothesis that all the observed accuracies come from chance. Denoting a the maximum accuracy of a given subject across TRs, the cor-

responding p-value was calculated as:

$$p^* = \frac{\sum_{i=1}^{N} I\left(\tilde{a}^{(i)} \geq a\right) + 1}{N + 1} \tag{3.6}$$

Both the Simes and maximum statistic correction are associated with the null hypothesis that no time-point has an effect, where the alternative is that at least one time-point has a true effect. Whereas the Simes correction is based on the use of dependent p-values associated with single time-points, the maximum statistic correction takes explicitly into account the correlation (under H_0) between different TRs, and therefore the results of the two procedures may vary slightly.

Second-level analysis was performed with a prevalence metric (Allefeld et al., 2016), suggested as a valid method to perform inference with cross-validated performance measures, both using the maximum permutation statistic p^* and the partial conjunction p-values $p^{u/m}$. The subject with the worst performance a_w, namely that subject with the highest p-value p_w, was employed to estimate the probability that such a value could be observed under the null hypothesis that the true effect has a prevalence γ lower than γ_0 in the population.

$$p(a_w|\gamma \leq \gamma_0) = [(1 - \gamma_0)\, p_w + \gamma_0]^{N_s} \tag{3.7}$$

where N_s denotes the number of subjects. As suggested in Allefeld et al. (2016), a sensible value for the prevalence is 50% of the population, which results in the following population p-value estimate:

$$p(a_w|\gamma \leq 0.5) = (0.5p_w + 0.5)^{N_s} \tag{3.8}$$

Results

Behavior

Listeners indicated not having difficulty segregating the stimulus from the continuous scanner noise due to its repetitive nature and a clear loudness difference. Subjects behavioral scores indicated that learning was successful after moderate training and that they could correctly identify targets at high performance levels in the aggregate (3.56 [3.12 4.05]; *Median edge-corrected d-prime [Inter Quartile Range]*)), bassoon (4.1 [3.54 4.05]), and cello (4.1 [3.39 4.05]) conditions. Observed edge-corrected d-prime values resulted from a decrease in subject's Hit rate alongside an increase in False Alarm rate, indicating that scores were not driven by a participant bias; furthermore, no bias was observed in false alarms which originated from trials that contained triplets in the unattended instrument as compared to those generated by trials with no triplets

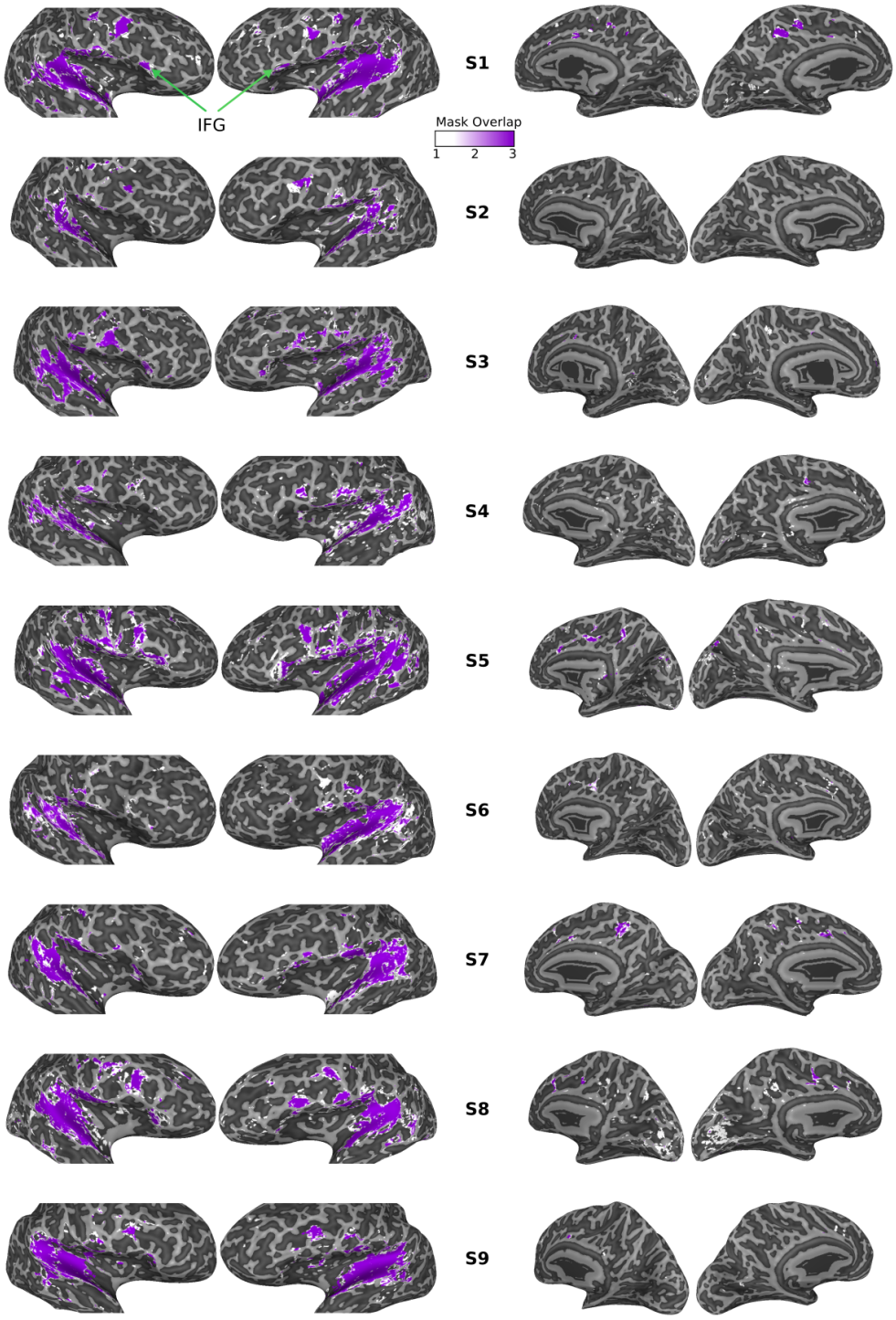


Figure 3.3: Subject IC-based ROIs. Temporal-frontal ICA-based ROIs per subject, showing the anatomical overlap between the three ROI masks employed in cross-validation (purple = all three masks overlap, white = single mask only; see Fig. 3.2); masks were independently generated from the training data for each cross-validation fold.

present. Accuracies across the different compositions were comparable at group level.

Temporal-frontal Network Estimation

We performed a ROI analysis based on ICAs, selecting those sound-responsive ICs containing a network of frontal-temporal cortical areas typically involved in attentive music listening. Such selection of ICs resulted in ROIs which were highly reliable within subjects and spatially comparable across subjects (Fig. 3.3); the number of voxels within the ROIs was left to vary freely across subjects (Fig. S3.1, left column). Temporal-frontal networks contained large sections of the Superior and Medial Temporal Gyrus (STG, MTG), including HG, PP, and PT, as well as sections of inferior parietal lobe including the Angular Gyrus (AG). Varying portions of medial and inferior frontal cortex were included, most notably the Inferior Frontal Gyrus (IFG). Inspection of mean responses across all voxels contained in the temporal-frontal networks showed similar shapes across subjects, displaying no overall pattern of mean difference between conditions (Fig. 3.4). Consistent with this observation, GLM-contrast maps showed no significant effects, neither during the full pre-triplet window (TR 1-8) nor per individual time-point.

Network-based Classification of Integration vs. Segregation trials

We further investigated a possible task effect within the functionally defined temporal-frontal ROI by virtue of a multivariate SVM-based decoding of integration versus segregation trials. After performing multiple-comparison correction based on the maximum permutation statistic p^*_m, decoding results showed significant above-chance classification at the $\alpha = .05$ level for at least one time-point in all subjects except S3 (Table S3.1). The Simes corrected p-values $p^{u/m}$, testing whether at least one of eight TRs showed an effect, resulted in significant effects at the $\alpha = .05$ level in all except subject S3 ($p^{u/m} = .0639$; Fig. 3.5a).

Second-level analysis via the prevalence metric, a measure suitable for group inference with cross-validated performance measures (see Multivariate Classification & Performance Evaluation), indicated a significant group effect both considering the maximum permutation statistic corrected p-value p^* ($\gamma_{p^*_m} = .0034$) and the partial conjunction p-value $p^{u/m}$ ($\gamma_{p^{u/m}} = .0031$). Confusion matrices for those time-points which passed multiple comparison correction showed no distinct confusion at group level (Fig. 3.5b), indicating that classifier performance did not result from a biased identification of single classes.

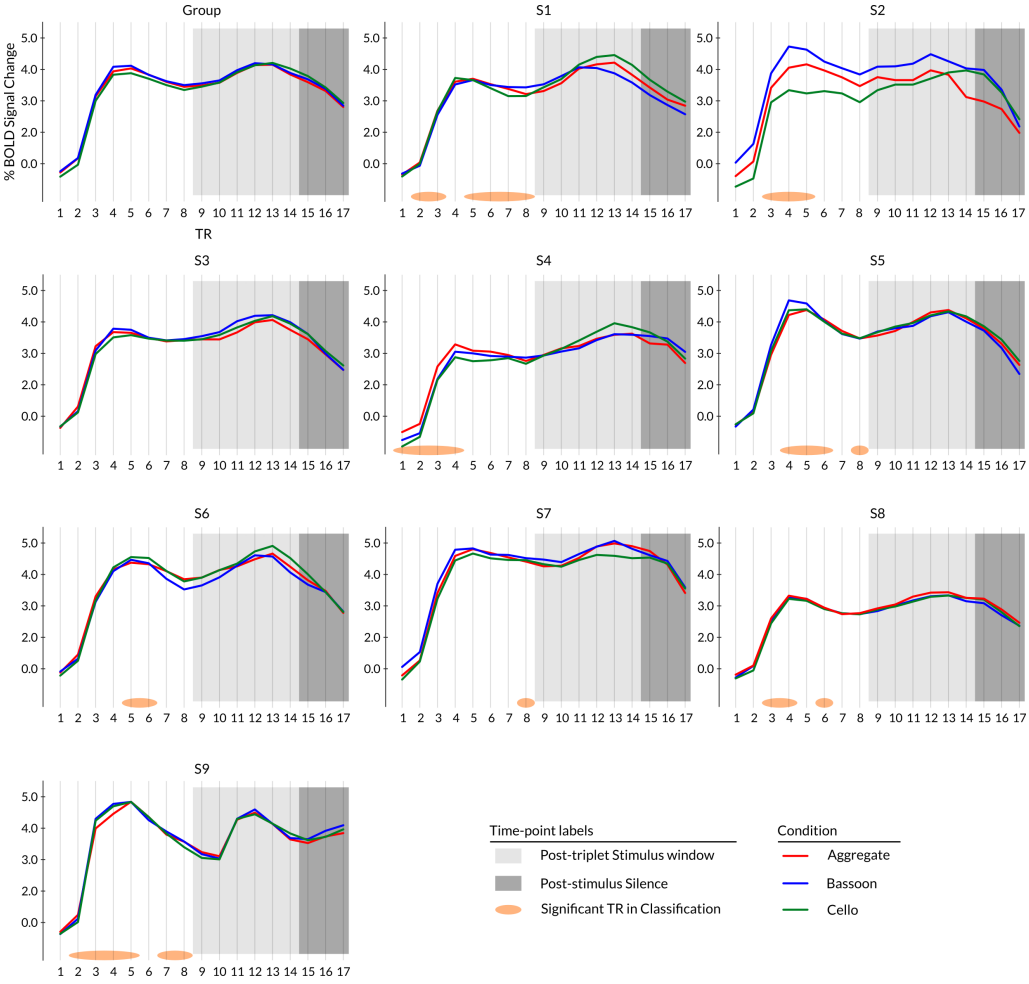


Figure 3.4: Voxel Responses. Mean responses per condition (% BOLD signal change) of all voxels included in the frontal-temporal ROI, for both the group (top-left) and individual subjects. Orange ellipses at the x-axis bottom indicate those pre-triplet time-points during which the integration versus segregation task could be classified significantly above chance after multiple comparison correction (see also Table S3.1). Red line, aggregate condition; blue line, bassoon condition; green line, cello condition; light-gray shading, post-triplet stimulus window; dark-gray shade, post-stimulus silence

ROI-based Classification of Integration vs. Segregation trials

To gain further insight into those spatial locations contributing to the distinction between integration and segregation conditions, we classified these conditions in an identical manner as for the temporal-frontal ROI, only now within the anatomically restricted temporal ROIs HG, PT, PP, anterior STG (aSTG), medial STG (mSTG), and posterior STG (pSTG; Fig. 3.6a) as well as for a difference ROI (Diff; Fig. S4) comprising all voxels not included within the temporal ROIs. All ROIs were individually defined for the left hemisphere (LH) and the right hemisphere (RH; see

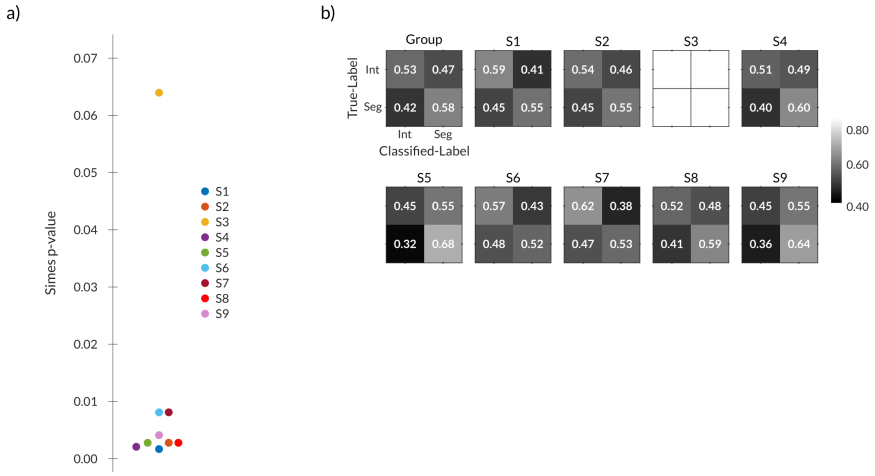


Figure 3.5: Network Classification Results. Integration versus segregation classification (a) as Simes *p*-values, testing whether at least one of subject's eight TRs showed an effect. Confusion matrices (b) for integration versus segregation classification, computed over all TRs which passed multiple comparison correction, for both the group (top left) and individual subjects.

Fig. S3.1 right-side for ROI sizes).

The Heschl's Gyrus ROI showed significant above-chance classification for all subjects, even though when inspecting results per hemisphere, this indicated no significant effects on the left for two subjects (S2 and S6) and only one on the right (S9; Fig. 3.6b). Medial STG (Fig. 3.6b) showed a pattern of higher classification in the RH, with two subjects (S2 and S8) not classifying significantly on the left and one on the right (S1) side. The difference ROI (Fig. 3.6b) showed highest accuracies in the LH, on both the left and the right one subject showed no significant classification (S3 and S2, respectively). Remaining areas showed similar patterns with above-chance classification in most subjects and ROIs (Fig. 3.6b & Fig. S3.3): PT (9/7, *number of subjects with left/right hemispheres showing above chance classification*), PP (9/8), aSTG (7/7), and pSTG (7/8). For all subjects except S4 there was at least one of the ROIs for which no time-points could be significantly classified (Fig. S3.3), as well as several ROI time-point combinations during which no subject showed significant classification. There does not appear to be a trend of specific time-points within subjects driving classification across ROIs, nor a clear underlying regional differentiation in average BOLD signal allowing for explanation of classifier performance differences (Fig. S3.4).

ROI-based Classification of Attended Instruments

We additionally tested whether the activation patterns within the ROIs, most notably in HG, enable the decoding of the attended instrument during the segregation trials, similar to the de-

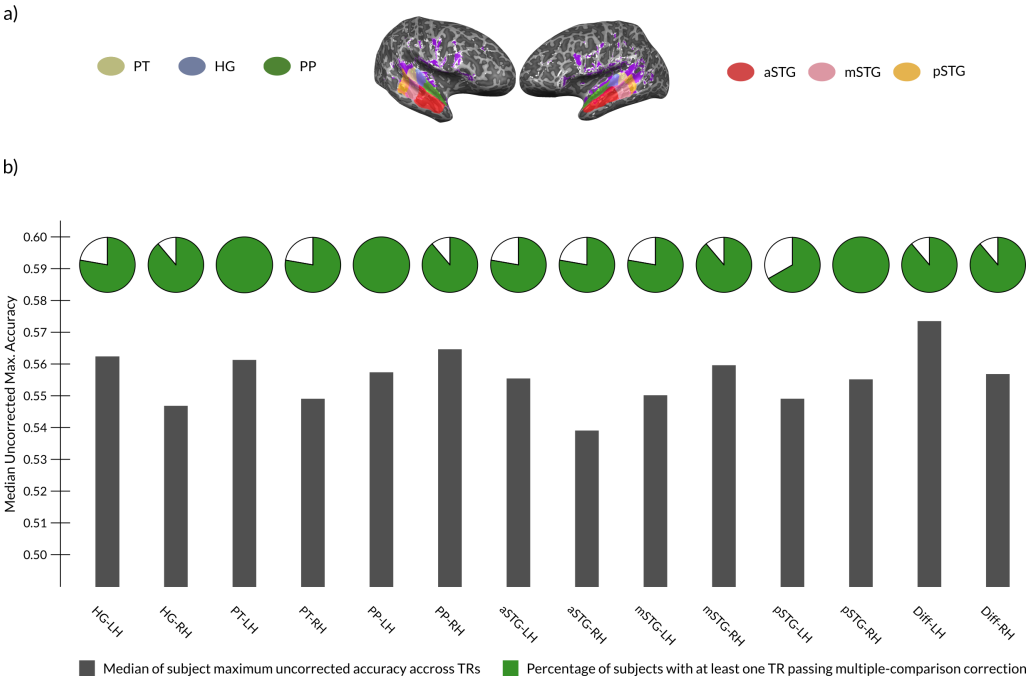


Figure 3.6: Temporal ROI Definition & Regional Classification Results. Example of left and right hemisphere temporal ROIs in a single subject (a); the Difference ROI (see Fig. S3.2) was defined as all voxels differing between the full temporal-frontal IC-based ROI (see Fig. 3.3) and the overlap of all temporal cortical anatomical ROIs (a). Classification results for integration versus segregation per ROI (b), displaying the group median of the highest uncorrected accuracy across a subject's TRs (gray bars) along with their respective pie-chart (green) indicating the percentage of subjects with at least one TR passing the multiple-comparison threshold.

coding of the attended speaker in classical speech-based paradigms. Multivariate classification with permutation-based inference was performed as described above, only now classifying attention to bassoon versus attention to cello trials. When inspecting the number of subject's with at least one TR passing multiple-comparison correction (Fig. 3.7, green pie-charts), HG showed significant above-chance classification at the $\alpha = .05$ level for all except subject S3. Inspection per hemisphere indicated a discrepancy between left and right, with the RH displaying significant classification in 8 out of 9 subjects while the LH only showed such effect in 4 listeners. The same trend can be observed in the HG's median of the maximum accuracy across subjects (Fig. 3.7, gray bars), where HG on the right is among the ROIs with the highest accuracy.

Discussion

In this work, we combined high-spatial resolution functional MRI at 7 Tesla with a previously validated psychophysical auditory scene analysis paradigm employing polyphonic music (see

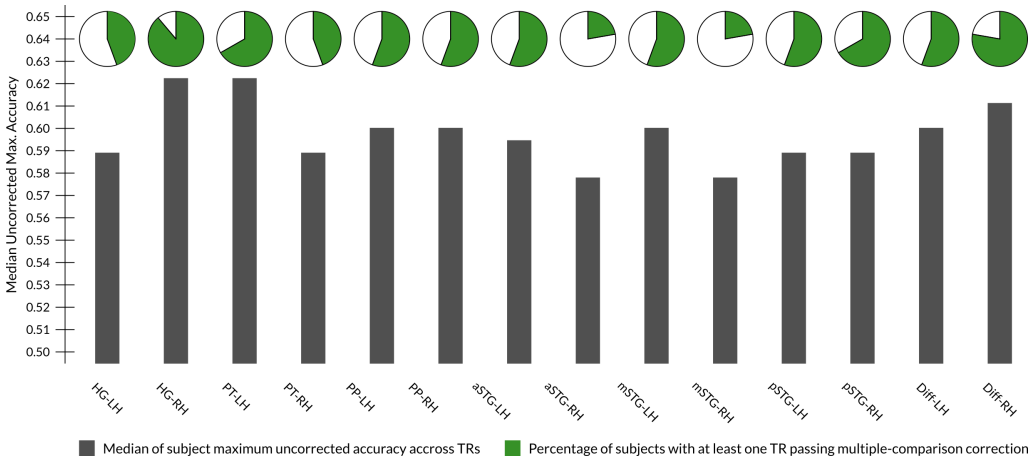


Figure 3.7: Classification Results for Bassoon Versus Cello. Displaying the group median of the highest uncorrected accuracy across a subject's TRs (gray bars) along with a pie-chart (green) indicating the percentage of subjects which had at least one TR passing the multiple-comparison threshold.

Disbergen et al., 2018, Chapter 2). During functional MRI measurements, subjects listened to polyphonic music consisting of two instruments, while their locus of attention was changed between individual instruments or the aggregate (*i.e.*, both instruments). Listeners were instructed to detect a triplet pattern located in the second half of the stimulus (Fig. 3.1).

Behavioral scores showed that subjects completed the task at desired high performance levels. Listener's triplet detection capacity indicated they managed to segregate the instruments presented in the mixture. If segregation had not been achieved, correct responses to the presence of triplets within single instruments would not have been possible, since unsegregated streams would only differ as regards their rhythmic cues (*i.e.*, tone on- and off-sets), based on which triplet assignment to either of the two voices would not have been possible. With this paradigm we aimed to investigate those cortical mechanisms allowing subjects to perceive melodic voices separately or integrate across them, even though the input arriving in their ear consisted solely of their mixed waveforms.

For the current investigation we focused on the contributions of each individual's (functionally defined) frontal-temporal network to attentive music listening (Fig. 3.3), specifically studying its modulation when subjects segregate or integrate musical voices. The results demonstrated that within the temporal-frontal ROI there were no significant effects present in the GLM-contrast maps, while attentional state could be classified above chance during the pre-target stimulus window on an individual subject basis (Fig. 3.5). To allow further specification of regional involvement within this network, we additionally investigated their contributions via anatomically defined ROIs, most notably showing that effects can be reliably obtained as early as in Heschl's

Gyrus (Fig. 3.6). Observed changes probably originated further up-stream from auditory cortex, driven by the cognitive differences which exist between the integration and segregation conditions and fed back to lower-level regions. The regional results within the Difference ROI indicated that both integration versus segregation and the individual segregation conditions can be reliably decoded within these non-auditory areas.

Distributed Attentional Network Involvement

Attentive conditions could be decoded across the multitude of ROIs included in the frontal-temporal IC (Fig. 3.7), suggesting that regions included in the fronto-temporal network are involved in both a purely cognitive (integration vs. segregation) and more physically (bassoon vs. cello) driven scene analysis task. Effects suggest that auditory areas form stimulus-driven representations which are modulated by attention, potentially originating in medial and inferior frontal cortical regions and displaying a flexible interaction between bottom-up and top-down driven mechanisms. Supramodal models of top-down control and attention have previously implicated extended front-parietal networks in similar tasks, which are typically separated into a dorsal and ventral branch (Corbetta & Shulman, 2002; Corbetta et al., 2008; Corbetta & Shulman, 2011). General auditory-based figure-ground segregation plays an important role in ASA and has been linked to these areas located outside of auditory cortex, more specifically the dorsal branch, implicating its involvement in a generalized top-down modulation of auditory cortical sound representations (e.g., Bizley & Cohen, 2013; Griffiths & Warren, 2004; Hausfeld et al., 2018; Shinn-Cunningham, 2008; Zatorre et al., 2002). Activity changes in the ventral branch of this prefrontal network have been linked more to the task-driven processing of auditory objects (Cohen et al., 2009; Hill & Miller, 2010), suggesting this hub may serve as the origin of the task-related modulations we observed here in (early) auditory cortex.

Regions typically comprised in our IC-based frontal-temporal network indeed show partial overlap with those classified as part of the ventral network, suggested as an amodal attention system (Corbetta & Shulman, 2002; Corbetta et al., 2008). The exact regional inclusion in the ventral network may be modality-specific and task-dependent, indicating the inclusion of posterior middle temporal gyrus and middle frontal gyrus as auditory-specific (Braga et al., 2013). In general, frontal cortex appears to play an important role in the higher-level cognitive processes which are an essential part of the ASA of complex sounds, allowing for the need of reliable and flexible representations of task-relevant sounds. Frontal areas could fulfill such role by virtue of downward projections of selective attention sources onto the auditory cortices, driving attention-based filtering in these areas which in turn provide feed-forward information assisting in the maintenance of representations along the cortical processing pipeline.

Region Specific Involvement

In the absence of physical differences between the presented stimuli, we observed top-down modulations of response patterns as early as in auditory cortex, including primary areas on HG. Previous studies have mostly focused on the segregation aspect of auditory scene analyses (e.g., Besle et al., 2011; Carlyon, 2003; Elhilali et al., 2009; Lakatos et al., 2013; Shamma & Micheyl, 2010; Sussman et al., 2007), while here we extended these observations into the music integration condition as well. We have demonstrated that activation patterns in HG still allowed for the decoding of the attended instrument during the segregation trials, implicating that HG activity is modulated by both bottom-up and top-down processing. During situations where the stimulus is acoustically identical, the analysis requirements for integration and segregation tasks differ strongly. When separating instruments, the neural populations representing the attended instrument, probably in a tonotopic fashion, could be enhanced while the unattended instrument is unaffected or suppressed (e.g., Fritz et al., 2007b,a; Lakatos et al., 2013; O'Connell et al., 2014; Da Costa et al., 2013). This is in line with previous work employing speech stimuli and demonstrating that the representations of talkers are modulated in early auditory cortex (e.g., Golumbic et al., 2013; Puschmann et al., 2018). An integrative task would require the attentional focus to be pooled more across the same single-instrument neural populations previously up- and/or down-regulated for segregation, hence creating neural conditions which would allow detection of activation pattern differences with fMRI.

The capacity of our models to differentiate between the segregation conditions within HG, provides an indication that early auditory areas contribute to observed effects and its activity is probably modulated by higher-order areas. Among the other temporal regions within which effects were detected, PT plays a role in general sound segregation processing by virtue of integrating both spectral-temporal and spatial information (Smith et al., 2010; Griffiths & Warren, 2002; Zatorre et al., 2002; Hausfeld et al., 2018). The PP and part of the anterior STG has been demonstrated to perform a more music-specific role in auditory processing, potentially in a right-lateralized fashion as well (Angulo-Perkins et al., 2014; Leaver & Rauschecker, 2010; Norman-Haignere et al., 2013). Within the frontal parts of the network, the IFG is of special interest due to its anatomical connections with both auditory belt and parabelt areas (Hackett, 2011; Kaas & Hackett, 2000; Rauschecker & Tian, 2000; Romanski & Averbeck, 2009), as well as its implication in generalized task-driven auditory processing (Atiani et al., 2014; Cohen et al., 2009; Hill & Miller, 2010). The angular gyrus has been hypothesized as a cross-modal attention-modulated combinatorial and integrative hub for multisensory information (for a review, see Seghier, 2012), potentially contributing to both music perception and production. Regional decomposition strengthens the observation that music ASA involves a combination of

regions which have been classically observed in scene analysis paradigms and music perception, operating alongside a larger-scale generalized top-down driven auditory attention network.

Limitations and Considerations

Attending to a single instrument in a musical mixture is a task which could potentially be more attentionally demanding than integrating the instruments, as typical music perception tends to be multi-instrumental. We do not observe behavioral differences between conditions, neither here nor in our larger-scale behavioral study (Disbergen et al., 2018, Chapter 2). However, this might be due to ceiling effects and/or partial insensitivity of the performance metric; see Disbergen et al. (2018, Chapter 2) for a more elaborate discussion. Importantly, we did not observe an overall change in BOLD response between conditions, as supported by the observation that GLM-contrast maps showed no significant effects, which would be expected in the case of large attentive load differences.

Even though we employed individually-defined temporal ROIs based on each subject's macroscopic anatomical landmarks, we cannot exclude that the HG-ROI contained non-primary auditory cortical tissue, aside from cautions based on the continuing discussion of the in vivo definition of primary AC (e.g., Baumann et al., 2013; Moerel et al., 2014). While we demonstrated that attentional modulations are reflected in large-scale cortical pattern changes, it is not possible to disambiguate at which time-points or delays these are present, nor which underlying neural changes may be driving the observed BOLD modulations.

Within the current work, we opted to acquire a large amount of functional and behavioral data per subject while limiting the number of participants as compared to typical sample sizes employed in human neuroimaging studies. Imaging data collection at ultra-high-field and high-resolution, provides improved sensitivity and specificity compared to the common lower fields, lower-resolution, and limited within-subject data typically employed in human neuroimaging (De Martino et al., 2018). Reduction of measurement error generally has a larger effect on model identifiability compared to sampling error reduction, *id est* more data versus more subjects, which appears to be independent of the signal to noise ratio (Kolossa & Kopp, 2018). In the current set-up we acquired on average 6 hours of imaging data for each subject, roughly equivalent to 3-4 subjects when employing common protocols. This increase in data quantity and quality is employed to allow for optimized and advanced within-subject analysis strategies, otherwise unfeasible with commonly acquired data-sets. Employing our analysis technique in combination with individual subject-level permutation-based statistical corrections, we have demonstrated significant effects in all of our participants alongside a significant group-level inference. Moreover, although there were some differences in the detail of individual cortical

activity patterns, there was a broad consistency across individuals which further supports the validity of our results.

Conclusion

By virtue of a novel within-subject multivariate classification-based analysis approach, this work has demonstrated that in music scene analysis, a large temporal-frontal network of sound-responding cortical areas is modulated by top-down attention. We showed that the task subjects perform - integration or segregation of the instruments - can be decoded above chance across most of the subject's individually defined frontal-temporal network. The more classically investigated sound segregation conditions (*i.e.*, attending a single instrument), also remain decodable above chance in a multitude of individual ROIs within this same network. Significant decoding of attentional state was observed as early as in HG, contrary to previous studies which have employed music stimuli. Current results extend the involvement of HG and its top-down modulation into the music realm, demonstrating effects similar to those previously shown for more schematic scene elements, as well as the special-case scenario of speech. Future research into the timing of these effects would allow insight into whether what we are observing here is related to the initial bottom-up driven analysis or modulations of ongoing sound-representations being fed back to lower-level auditory areas.

References

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1515), 285–299.

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *Neuroimage*, 141(C), 378–392.

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage*, 20(2), 870–888.

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., & Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. *CORTEX*, 59(C), 126–137.

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent Selectivity for Task-Relevant Stimuli in Higher-Order Auditory Cortex. *Neuron*, 82(2), 486–499.

Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Task Difficulty and Performance Induce Diverse Adaptive Patterns in Gain and Shape of Primary Auditory Cortical Receptive Fields. *Neuron*, 61(3), 467–480.

Baumann, S., Petkov, C. I., & Griffiths, T. D. (2013). A unified framework for the organization of the primate auditory cortex. *Frontiers in Systems Neuroscience*, 7.

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Emerson, R. G., & Schroeder, C. E. (2011). Tuning of the Human Neocortex to the Temporal Dynamics of Attended Events. *The Journal of Neuroscience*, 31(9), 3176–3185.

Bigand, E., Foret, S., & McAdams, S. (2000). Divided attention in music. *International Journal of Psychology*, 35(6), 270–278.

Bizley, J. K. & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10), 693–707.

Braga, R. M., Wilson, L. R., Sharp, D. J., Wise, R. J. S., & Leech, R. (2013). Separable networks for top-down attention to auditory non-spatial and visuospatial modalities. *Neuroimage*, 74(C), 77–86.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The Perceptual Organization of Sound. Cambridge, Massachusetts: MIT Press.

Bregman, A. S. & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1), 19–31.

Brochard, R., Drake, C., Botte, M. C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1742–1759.

Carlyon, R. P. (2003). How the brain separates sounds. *Trends in Cognitive Sciences*, 8(10), 465–471.

Carlyon, R. P. & Cusack, R. (2005). Effects of Attention on Auditory Perceptual Organization. In L. Itti, G. Rees, & J. K. Tsotos (Eds.), *Neurobiology of Attention* (pp. 317–323). Cambrige, MA: Elsevier.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research*, 352, 49–69.

Coffey, E. B. J., Scala, S., & Zatorre, R. J. (2011). Montreal Music History Questionnaire: a tool for the assessment of music-related experience. In *Neurosciences and Music IV Learning and Memory* Edinburgh, UK.

Cohen, Y. E., Russ, B. E., Davis, S. J., Baker, A. E., Ackelson, A. L., & Nitecki, R. (2009). A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. *Proceedings of the National Academy of Sciences*, 106(47), 20045–20050.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, 58(3), 306–324.

Corbetta, M. & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215.

Corbetta, M. & Shulman, G. L. (2011). Spatial Neglect and Attention Networks. *Annual review of neuroscience*, 34(1), 569–599.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656.

Cusack, R. & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, 62(5), 1112–1120.

Da Costa, S., van der Zwaag, W., Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning In to Sound: Frequency-Selective Attentional Filter in Human Primary Auditory Cortex. *The Journal of Neuroscience*, 33(5), 1858–1863.

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludağ, K., De Weerd, P., Ugurbil, K., Goebel, R., & Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. *Neuroimage*, 168, 366–382.

Deutsch, D. (2013). Grouping Mechanisms in Music. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 183–248). London, UK: Elsevier.

Ding, N. & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29), 11854–11859.

Disbergen, N. R., Valente, G., Formisano, E., & Zatorre, R. J. (2018). Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in Neuroscience*, 12.

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. *PLoS Biology*, 7(6), 1–14.

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., Miller, K. L., Ugurbil, K., & Yacoub, E. (2010). Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging. *PLoS ONE*, 5(12).

Formisano, E., Esposito, F., Di Salle, F., & Goebel, R. (2004). Cortex-based independent component analysis of fMRI time series. *Magnetic Resonance Imaging*, 22(10), 1493–1504.

Fritz, J., Shamma, S. A., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, 6(11), 1216–1223.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007a). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hearing Research*, 229(1-2), 186–203.

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007b). Adaptive Changes in Cortical Receptive Fields Induced by Attention to Complex Sounds. *Journal of Neurophysiology*, 98(4), 2337–2346.

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., & Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *Neuroimage*, 132, 32–42.

Golland, P., Liang, F., Mukherjee, S., & Panchenko, D. (2005). Permutation Tests for Classification. In *Machine Learning and Interpretation in Neuroimaging: International Workshop, MLINI 2011, Held at NIPS 2011, Sierra Nevada, Spain, December 16-17, 2011 …* (pp. 501–515). Berlin, Heidelberg: Springer Berlin Heidelberg.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party". *Neuron*, 77(5), 980–991.

Gregory, A. H. (1990). Listening to Polyphonic Music. *Psychology of Music*, 18(2), 163–170.

Griffiths, T. D. & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, 25(7), 348–353.

Griffiths, T. D. & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5, 887–892.

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hearing Research*, 271(1-2), 133–146.

Hausfeld, L., Riecke, L., & Formisano, E. (2018). Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. *Neuroimage*.

Heller, R., Golland, Y., Malach, R., & Benjamini, Y. (2007). Conjunction group analysis: An alternative to mixed/random effect analysis. *Neuroimage*, 37(4), 1178–1185.

Hill, K. T. & Miller, L. M. (2010). Auditory Attentional Control and Selection during Cocktail Party Listening. *Cerebral Cortex*, 20(3), 583–590.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3), 626–634.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective, & Behavioral Neuroscience*, 2(2), 121–140.

Jones, S. E., Buchbinder, B. R., & Aharon, I. (2000). Three-dimensional mapping of cortical thickness using Laplace's Equation. *Human Brain Mapping*, 11(1), 12–32.

Kaas, J. H. & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences*, 97(22), 11793–11799.

Kim, J. J., Crespo-Facorro, B., Andreasen, N. C., O'Leary, D. S., Zhang, B., Harris, G., & Magnotta, V. A. (2000). An MRI-Based Parcellation Method for the Temporal Lobe. *Neuroimage*, 11(4), 271–288.

Kolossa, A. & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *Neuroimage*, 172, 775–785.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, 77(4), 750–761.

Leaver, A. M. & Rauschecker, J. P. (2010). Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *The Journal of Neuroscience*, 30(22), 7604–7612.

Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, 30(3), 259–274.

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., & Gruetter, R. (2010). MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage*, 49(2), 1271–1281.

McAdams, S. (2013a). Musical timbre perception. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 35–68). London, UK: Elsevier Inc.

McAdams, S. (2013b). Timbre as a structuring force in music. In *ICA 2013 Montreal* (pp. 1–6).: ASA.

McDermott, J. H. & Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology*, 18(4), 452–463.

McKeown, M. (2003). Independent component analysis of functional MRI: what is signal and what is noise? *Current Opinion in Neurobiology*, 13(5), 620–629.

Mesgarani, N. & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397), 233–236.

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Ugurbil, K. (2010). Multiband Multislice GE-EPI at 7 Tesla, With 16-Fold Acceleration Using Partial Parallel Imaging With Application to High Spatial and Temporal Whole-Brain FMRI. *Magnetic Resonance in Medicine*, 63(5), 1144–1153.

Moerel, M., De Martino, F., & Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8.

Nichols, T. E. & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 1–25.

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical Pitch Regions in Humans Respond Primarily to Resolved Harmonics and Are Located in Specific Tonotopic Regions of Anterior Auditory Cortex. *The Journal of Neuroscience*, 33(50), 19451–19469.

O'Connell, M. N., Barczak, A., Schroeder, C. E., & Lakatos, P. (2014). Layer Specific Sharpening of Frequency Tuning by Selective Attention in Primary Auditory Cortex. *The Journal of Neuroscience*, 34(49), 16496–16508.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7), 1697–1706.

Paltoglou, A. E., Sumner, C. J., & Hall, D. A. (2009). Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. *Hearing Research*, 257(1-2), 106–118.

Peretz, I. & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology*, 56(1), 89–114.

Pressnitzer, D., Suied, C., & Shamma, S. A. (2011). Auditory scene analysis: the sweet music of ambiguity. *Frontiers in Human Neuroscience*, 5, 1–11.

Puschmann, S., Baillet, S., & Zatorre, R. J. (2018). Musicians at the Cocktail Party: Neural Substrates of Musical Training During Selective Listening in Multispeaker Situations. *Cerebral Cortex*, 1537, 224–13.

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and Integration of Auditory Streams when Listening to Multi-Part Music. *PLoS ONE*, 9(1), 1–9.

Rauschecker, J. P. & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22), 11800–11806.

Riecke, L., Peters, J. C., Valente, G., Kemper, V. G., Formisano, E., & Sorger, B. (2016). Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cerebral Cortex*, advance online access, 1–13.

Romanski, L. M. & Averbeck, B. B. (2009). The Primate Cortical Auditory System and Neural Representation of Conspecific Vocalizations. *Annual review of neuroscience*, 32(1), 315–346.

Seghier, M. L. (2012). The Angular Gyrus. *The Neuroscientist*, 19(1), 43–61.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5), 1210–1224.

Shamma, S. A. & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3), 361–366.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5), 182–186.

Smith, K. R., Hsieh, I.-H., Saberi, K., & Hickok, G. (2010). Auditory Spatial and Object Processing in the Human Planum Temporale: No Evidence for Selectivity. *Journal of Cognitive Neuroscience*, 22(4), 632–639.

Smith, S. M. & Brady, J. M. (1997). *SUSAN - a new approach to low level image processing*. *International Journal of Computer Vision*, 23(1), 45–78.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23(S1), 208–219.

Snyder, J. S. & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin*, 133(5), 780–799.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137–149.

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the acoustical society of America*, 117(3), 1285–14.

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, 69(1), 136–152.

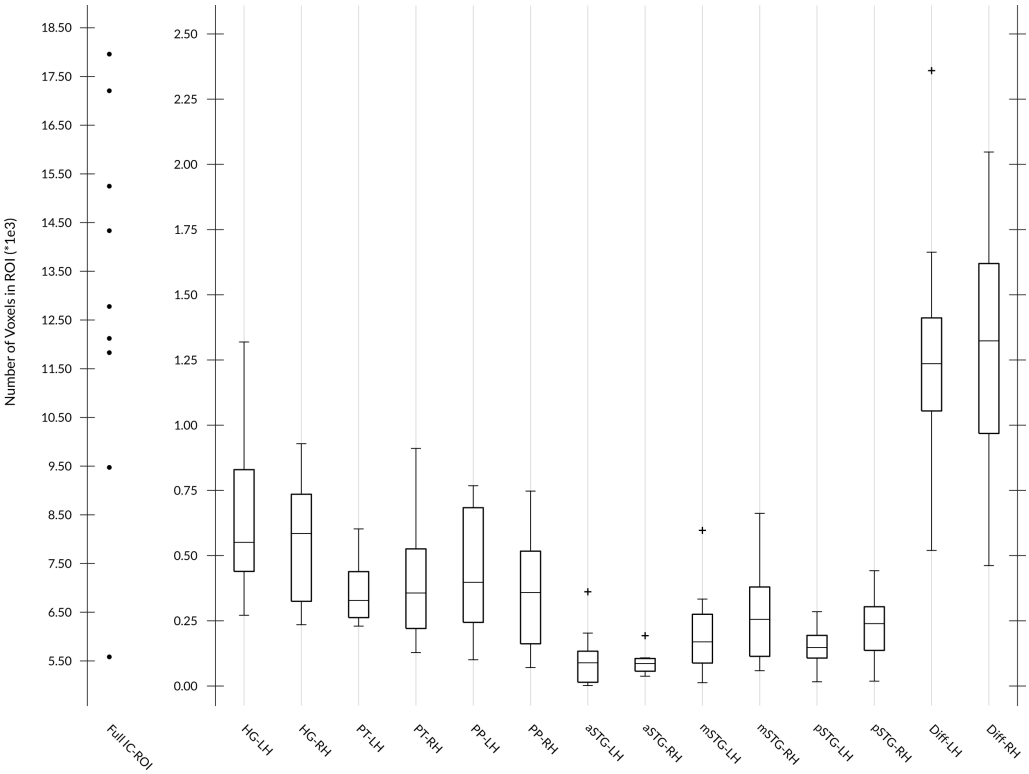
Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *Neuroimage*, 77, 52–61.

Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2), 45–52.

Zatorre, R. J., Bouffard, M., Ahad, P., & Belin, P. (2002). Where is 'where' in the human auditory cortex? *Nature Neuroscience*, 5(9), 905–909.

Zatorre, R. J. & Zarate, J. M. (2012). Cortical Processing of Music. In *The Human Auditory Cortex* (pp. 261–294). New York, NY: Springer New York.

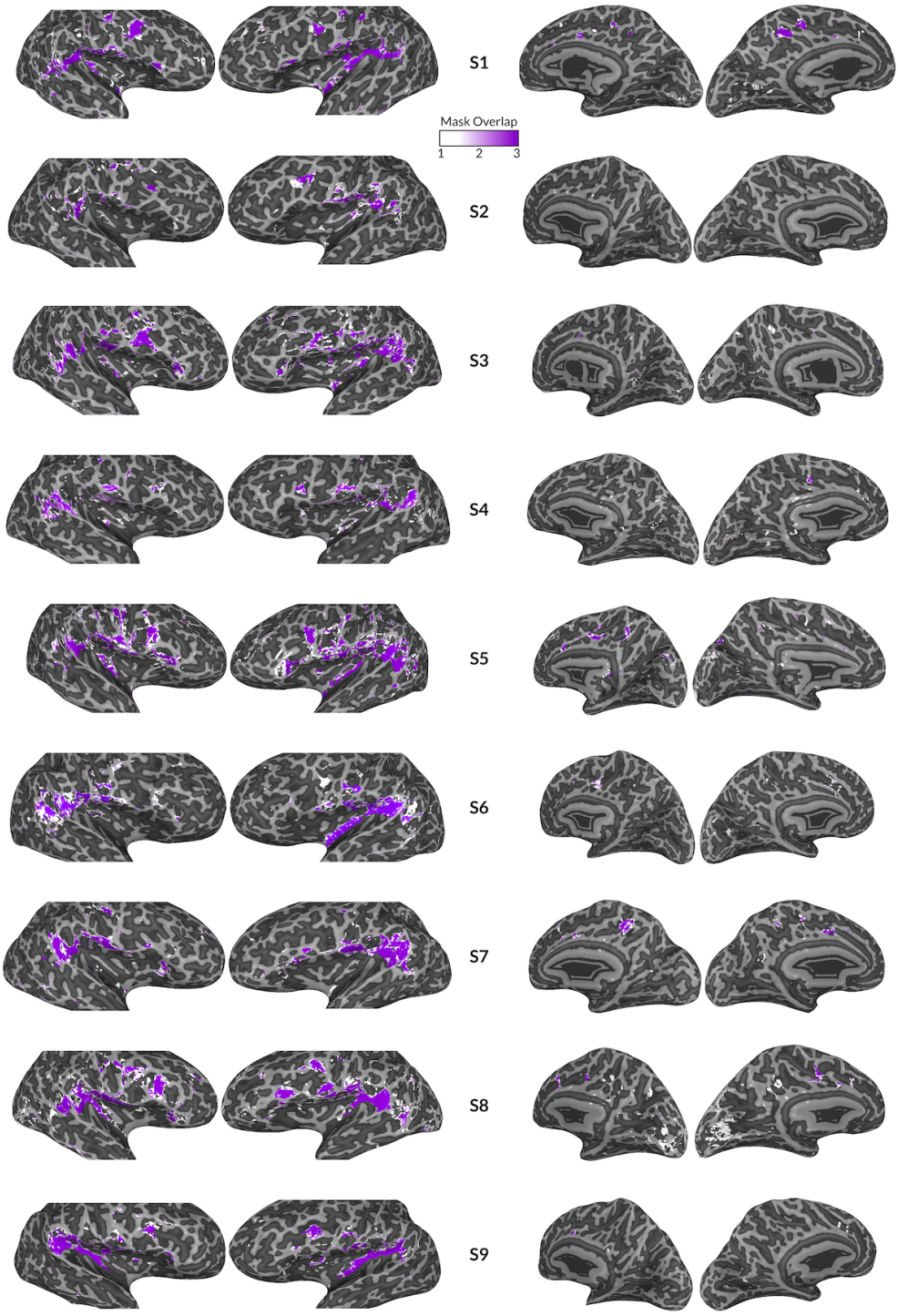
Supplementary Figures



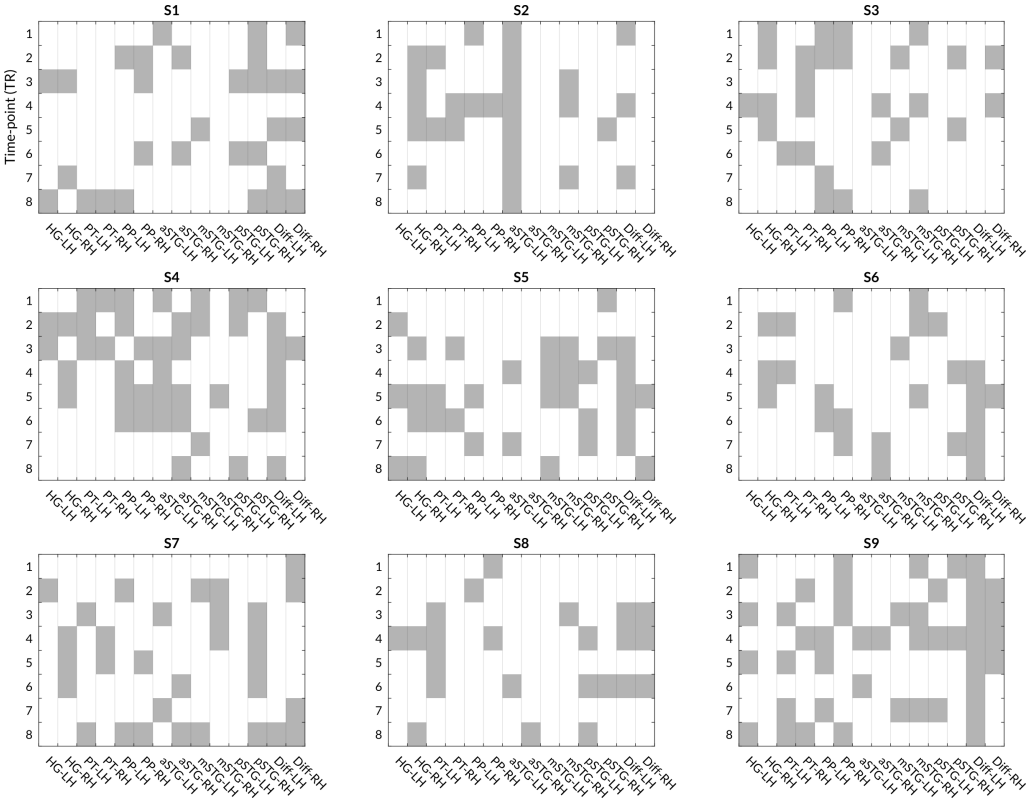
Supplementary Figure 3.1: ROI Sizes. Number of voxels included in the full temporal-frontal network (left-most) and all the ROIs (box = 25th percentile - median - 75th percentile; see Fig. 3.6 for ROIs.)

	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total
TR1	0.875	1.000	0.482	0.001	1.000	1.000	1.000	0.875	0.848	1
TR2	0.004	1.000	0.051	0.001	1.000	1.000	0.151	1.000	0.014	3
TR3	0.001	0.001	1.000	0.002	0.639	0.827	0.145	0.001	0.029	5
TR4	0.174	0.001	0.114	0.002	0.001	0.339	1.000	0.001	0.029	4
TR5	0.001	0.001	0.999	1.000	0.001	0.001	0.997	0.494	0.018	5
TR6	0.001	0.991	1.000	1.000	0.018	0.009	1.000	0.001	0.185	4
TR7	0.001	0.776	1.000	0.324	0.714	1.000	1.000	1.000	0.001	2
TR8	0.001	1.000	0.993	1.000	0.001	1.000	0.001	1.000	0.001	4
Total	6	3	0	4	3	2	1	3	6	-

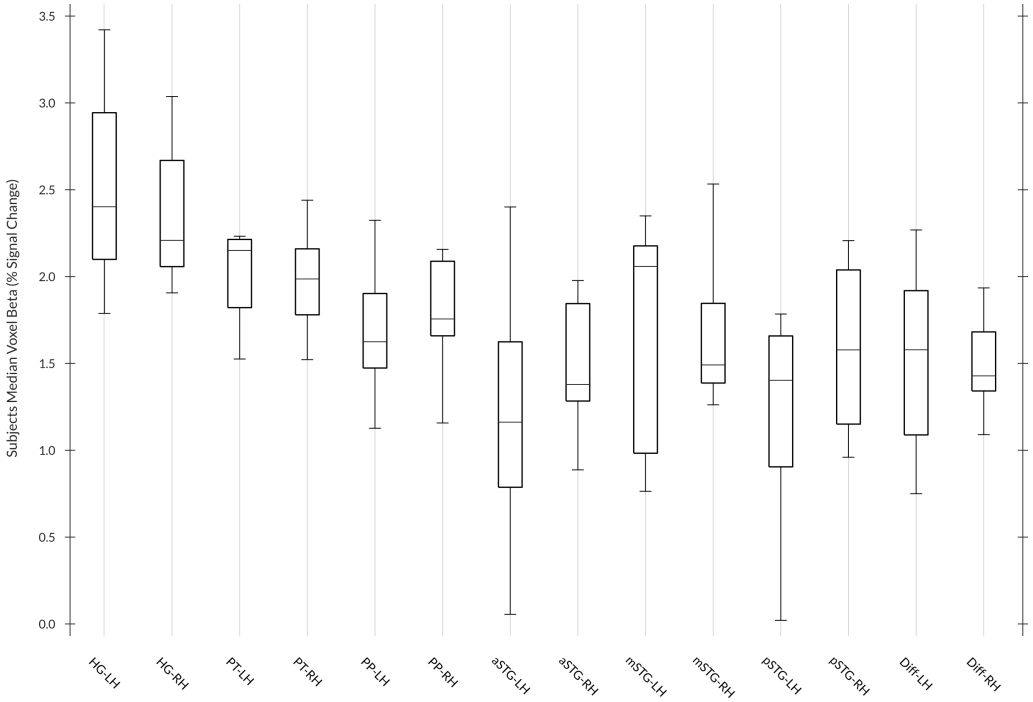
Supplementary Table 3.1: Network Classification Results. Maximum permutation statistic corrected p-values (p^*_m) for each time-point and subject, indicating those TRs passing the $\alpha = .05$ threshold with gray boxes, as well as the total number of subjects/time-points meeting this criterion.



Supplementary Figure 3.2: Difference ROIs. Displaying the Difference ROI overlap per subject. Regions include all those voxels included in the frontal-temporal ROIs (Fig. 3.3) which are not part of any anatomical temporal-ROIs (Fig. 3.6a).



Supplementary Figure 3.3: Regional Classification Overview. Per subject integration versus segregation classification accuracies passing multiple comparison correction (gray), separated for each ROI and time-point.



Supplementary Figure 3.4: Group ROI Signals. Group median Betas (in % signal change) per ROI across all voxels and trials (box = 25th percentile - median - 75th percentile).

"All you really need to know for the moment is that the universe is a lot more complicated than you might think, even if you start from a position of thinking it's pretty damn complicated in the first place"

Douglas Noel Adams

4

Modulating cortical instrument-representations during auditory stream segregation and integration with polyphonic music

Based on: Disbergen, N. R.†, Hausfeld, L.†, Valente, G., Zatorre, R. J., and Formisano, E. (to be submitted). Modulating cortical instrument-representations during auditory stream segregation and integration with polyphonic music. † equal contribution.

Abstract

Music containing multiple instruments can be appreciated by focusing on all instruments simultaneously (*i.e.*, integration) or on its individual entities (*i.e.*, segregation). Here we investigated the neural correlates of attentive listening to segregated or integrated music instruments using electroencephalography (EEG) and sound envelope reconstruction methods. We made use of twenty uniquely composed music pieces played by bassoon and cello and a previously validated behavioral paradigm for music auditory scene analysis (Disbergen et al., 2018, Chapter 2). Training both single- and multi-delay EEG-based decoders and examining their capacity to reconstruct the music envelopes. During the segregation task, attended instruments could be reconstructed better than unattended ones. This effect was present during a middle-latency window for both the bassoon (170-270ms) and cello (150-200ms). In addition, an attention effect was found during a late window (320-370 ms) only when reconstructing the bassoon envelopes. During the integration task, neither delay-general nor delay-resolved models displayed significant attentive modulations to the music envelopes. Subsequent analyses indicated that this may be due to heterogeneous strategies listeners employ during the integration task. Results from the segregation task suggest that for polyphonic music scenes, the segregation of an individual instrument (*i.e.*, stream) is achieved through the top-down modulation of the relevant instrument's neuronal representation, which occurs subsequent to an initial acoustically-driven scene segmentation.

Introduction

Listening to a sound of interest in an environment with multiple competing sounds represents a common though challenging task which the auditory system solves seemingly without effort. When this sound of interest is music, our auditory system is able to segregate it into its individual components (or streams) which represent, for example, multiple simultaneously playing instruments. The perceptual mechanisms for analyzing and resolving auditory (and musical) scenes have been described in a comprehensive theoretical framework by Bregman (1990). Research inspired by Bregman's theory has detailed the conditions under which acoustical scene elements are segregated or integrated, driven by physical differences between sounds (*i.e.*, bottom-up) as well as by top-down mechanisms, among which the listener's locus of attention (Besle et al., 2011; Bregman, 1990; Brochard et al., 1999; Carlyon, 2003; Carlyon & Cusack, 2005; Cusack et al., 2004; Lakatos et al., 2013; Shamma & Micheyl, 2010; Riecke et al., 2016; Sussman et al., 2007). Here we focus on the contributions of top-down attentive processes to auditory scene analysis (ASA) in the context of music listening, even though other cues, such as perceptual modulations by previous exposure, form an essential part of top-down mechanisms as well (e.g., Bey & McAdams, 2003; McAdams & Bregman, 1979; Bregman, 1990).

Most studies investigating ASA mechanisms have employed simple auditory scenes such as pure tones in noise or alternating tone sequences (for reviews see, Alain & Bernstein, 2015; Bregman, 1990, 2015; Carlyon, 2003; Ciocca, 2008). Since the auditory system has been optimized to process sounds that are relevant for behavior, naturalistic auditory scenes with ecologically valid stimuli are very valuable to gain a better understanding of ASA (for a review, see Theunissen & Elie, 2014). To date, most research on ASA with naturalistic stimuli has focused on language and employed multi-speaker environments in combination with selective attention tasks. Several studies used these paradigms in conjunction with magnetoencephalography (MEG), electroencephalography (EEG), or electro-corticography (ECoG) and identified effects of selective attention using sound envelope tracking/reconstruction methods (e.g., Crosse et al., 2015; Dijkstra et al., 2015; Ding & Simon, 2012b; Kerlin et al., 2010; Kubanek et al., 2013; Nourski et al., 2009; O'Sullivan et al., 2015). This research showed that for scenes containing two simultaneous speakers, the attended speech could be better reconstructed as compared to the unattended speech (Ding & Simon, 2012b,a; Mirkovic et al., 2015) at delays of approximately 100ms onwards (Hausfeld et al., 2018; O'Sullivan et al., 2015; Power et al., 2012). This suggests an attention-mediated biasing mechanism which enhances the neural representation of the relevant speech stream after an initial acoustical analysis of the sound mixture.

The investigation of multi-speaker scenes has provided insight into the processing of speech,

however a generalization of these mechanisms to auditory scenes including sounds other than speech is not straight-forward and requires additional investigations (e.g., Alho et al., 2014). Music, especially when containing multiple instruments (*i.e.*, polyphonic), is very well suited for the investigation of ASA in naturalistic and complex listening scenarios. Polyphonic music contains rich but acoustically well-controlled sound mixtures with a continuously varying degree of spectral and temporal overlap. Furthermore, multi-instrument music allows for the study of both the typical segregation aspect of ASA as well as the less investigated integration condition.

Schaefer et al. (2011) demonstrated, in musically experienced participants, that there is a high correlation between the evoked response potentials (ERPs) and the envelopes of the musical stimuli, mostly at 70-100ms after stimulus onset. They proposed that these correlations are representative of bottom-up processing potentially occurring outside the focus of attention. Treder et al. (2014) reported similar effects, even though during later times, around 200ms post-stimulus onset. They compared ERP responses for attended and unattended instruments within multi-instrument music which contained standard or deviant structures in the individual instruments. Results suggest that higher-level cortical processing influenced the ongoing sound representations, specifically of the to-be-attended instrument. Taken together, these studies indicate that music envelopes are represented in the EEG signal and are, similarly to speech, modulated by attention during later time-windows. These studies should be interpreted with caution since investigations of music stream representation and attentive modulation have mostly focused on expert musicians, who typically display modified listening behavior as compared to non-musicians (e.g., Puschmann et al., 2018; Coffey et al., 2017). Very few studies have investigated the processes involved in auditory stream integration, and even less have used music stimuli in their investigation (Deutsch, 2013; Disbergen et al., 2018; Ragert et al., 2014; Sussman, 2005; Uhlig et al., 2013).

In our functional magnetic resonances imaging (fMRI) study employing a music ASA paradigm (Chapter 3), we demonstrated that integrating or segregating music instruments, as well as focusing attention on only one of the instruments, resulted in differential cortical activity patterns in a large frontal-temporal network of sound-responding cortical regions. This network included several regions early in the auditory processing hierarchy, such as Heschl's gyrus (HG). Even though fMRI is well suited to localize the effects of attention, it is less well suited to determine the time-course and order of effects. For example, results in HG could have originated from both an early modulation of the initial bottom-up driven sound analysis as well as later top-down driven mechanisms that influence sustained responses in HG through feedback. Here we aimed to investigate these same attention effects with a high temporal resolution and identify when these attentive effects take place. To this end, we employed a previously validated psycho-physical paradigm (Disbergen et al., 2018, Chapter 2) in combination with an EEG envelope-based neural

tracking method (Hausfeld et al., 2018). Non-musicians performed a task requiring stream segregation or integration of custom-composed polyphonic music pieces, attending either single instruments or integrating across them.

During the segregation condition, we expected higher reconstruction accuracy when an instrument was attended to as opposed to unattended. Effects were predicted at delays beyond 100ms due to earlier windows mostly representing initial bottom-up mechanisms, driving the stimulus processing based on acoustical features. Effects during early delay windows could additionally contain preparatory top-down influences, the segregation of which from bottom-up mechanisms is not possible with the current design. In general, early windows are not expected to be strongly biased by attentive mechanism, albeit subtle modulations may already take place (e.g., Poghosyan & Ioannides, 2008; Woldorff & Hillyard, 1991). Integration of instruments was hypothesized to differ from segregation mostly regarding its timing, since integration is typically understood as a cognitively higher-level task as compared to segregation, hence potential differences are expected to emerge during later delay-windows for instrument integration as compared to segregation.

Methods

Subjects

Nineteen adult volunteers (10 women; age 23.9 +/- 3.3 years, *mean +/- standard deviation*) with self-reported normal motor and vision abilities participated in this study. All subjects displayed normal hearing thresholds (<25 decibels Hearing Level), as measured via pure-tone audiometry in both ears at frequencies of 0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, and 6.0 kHz. None of the participants spoke a tonal language and all had less than two years of (formal) musical training on a lifetime basis with instruments which were not bassoon or cello, as assessed via the Montreal Music History Questionnaire (Coffey et al., 2011). Volunteers were mostly students recruited from Maastricht University and provided written informed consent in accordance with the protocol as approved by the Maastricht University Ethics Review Committee Psychology and Neuroscience (#167_09_05_2016). Five subjects were excluded from the EEG analysis due to low behavioral performance metrics in one or multiple conditions, hence subsequent analyses were performed on 14 participants (see also Fig. S4.1).

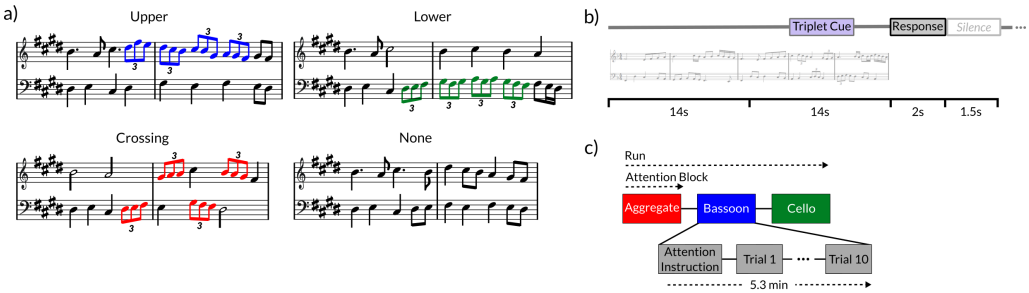


Figure 4.1: Experiment Design. Different triplet versions for each music composition (a): upper voice (*i.e.* bassoon; blue notes), lower voice (*i.e.* cello; green notes), crossing voices (red notes), no triplets. Trial buildup (b) with 28s stimulus, 2s response window, and 1.5s silence. Trials were presented in attentive blocks of 10 stimuli each and preceded by a visual attention instruction and silence.

Stimuli

In this experiment we employed a previously validated psychophysical paradigm for the study of ASA with multi-instrument music. For the reader's convenience, we include an overview below while an in-depth discussion of the stimuli, task, training, and a demonstration of its validity, including its use with non-musically trained subjects, can be found in Disbergen et al. (2018, Chapter 2). Twenty custom-composed polyphonic counterpoint music pieces (28s duration) consisting of two instrument voices were synthesized for bassoon (treble clef) and cello (bass clef) at a tempo of 60 beats per minute. Melodies were synthesized independently for bassoon and cello from Musical Instrument Digital Interface (MIDI) files, with a sampling rate of 44.1 kHz and a 16 Bits resolution in Logic Pro 9 (Apple Inc., Cupertino, California, USA). Resulting stimuli were post-hoc combined into polyphonic pieces with Root Mean Square (RMS) equalization and their onsets and offsets exponentially ramped with a rise-fall time of 100ms. All stimulus processing and manipulation aside from synthesizing was performed with custom-developed MATLAB codes (The MathWorks Inc., Natick, Massachusetts, USA).

We aimed to examine modulations of a musical instrument's neural representations by both comparing the integration versus segregation listening conditions as well as attended versus unattended within the segregation conditions. To achieve these different listening contexts under fixed acoustic circumstances, we varied the listener's focus of attention using a temporal detection task, which was implemented through rhythmic modulations incorporated in the polyphonic music (see Disbergen et al., 2018, Chapter 2). Modulations comprised four consecutive triplets (total duration 4s), each containing three eighth notes played in one single beat and carefully integrated into the melodic structure (see Fig. 4.1a). Patterns of four consecutive triplets could be located in the upper voice melody (*i.e.*, bassoon; Fig. 4.1a, blue notes), lower voice (*i.e.*, cello; Fig. 4.1a, green notes), across voices (Fig. 4.1a, red notes), or not be present. If triplets were located across instruments, they started randomly in bassoon or cello and alternated voices ac-

cordingly, while patterns present in a single voice were only located within that respective instrument. Triplets were always incorporated in the second half of the melodies, pseudo-randomly starting between 14 to 19s after music onset, resulting in stimuli which were physically identical up until triplet occurrence. Temporal modulations in the form of triplets were chosen due to their orthogonality towards pitch based segregation mechanisms, facilitating their detection by listeners with little to no musical training.

Paradigm

Listeners were instructed to complete a forced-choice delayed-response target detection task within or across instruments, attending the same instrument(s) during an attention block of 10 consecutive trials (Fig. 4.1c). Each trial comprised the music stimulus of 28s, a 2s response window, and a 1.5s silence (Fig. 4.1b). A visual instruction was presented before the beginning of each attention block, cuing which instrument(s) to attend: bassoon, cello, or aggregate (*i.e.*, both instruments; Fig. 4.1c). After the stimulus ended, listeners responded via a button-press whether the triplet pattern was present in those instrument(s) they were instructed to attend. Stimuli were presented pseudo-randomly in sets of three consecutive attention blocks of 10 trials each, covering all three attention conditions. This three-block scheme was repeated four times, covering all stimuli under all attention conditions twice, hence resulting in two fully balanced experiment repetitions. Each attention block of 10 trials contained five control-trials with triplets located either in the unattended instrument or no triplets present. For the bassoon and cello conditions, the inclusion of five control-trials resulted in an uneven distribution of no-triplet and opposite-to-attention-voice trials. To mitigate possible effects caused by this imbalance, the number of trials of each version was alternated between experiment repetitions, three-four or four-three. When attending to both instruments, only control trials containing no triplets were employed.

Due to the limited musical education of participants, they were first subjected to a training session during which they listened to music slowly increasing in complexity, initiating with scales including individual triplets and completing with melodies containing the triplet patterns and at equal complexity as the actual experiment; for training details see Disbergen et al. (2018, Chapter 2).

EEG Data Acquisition & Preprocessing

Electroencephalographic scalp-data was recorded in an electrical and sound insulated chamber from 63 electrodes using BrainAmp amplifiers (Brain Products, Munich, Germany) in a modified 10-20% electrode system (EasyCap, montage 11) and referenced to electrode TP9. The vertical

and horizontal electro-oculogram (EOG) were recorded from electrodes placed below and next to the right eye. During acquisition, the electrodes' impedance was kept below 5$k\Omega$, the EEG signal was bandpass-filtered with an analog filter at cutoffs 0.01 and 200Hz and digitized at a 500Hz sampling rate. EEG data preprocessing was performed using the EEGLAB toolbox (Delorme & Makeig, 2004) in MATLAB and custom MATLAB codes. Preprocessing steps included band-pass filtering with a finite impulse response (FIR) filter at cutoffs 0.5 and 45Hz, re-referencing to an average electrode reference, and epoching from 1-28s relative to the onset of the auditory stimulus. Epoched data was fed to an independent component analysis (ICA) for artifact removal, employing the EEGLAB *runica()* function. Component estimation was followed by a manual definition of artifact components containing eye movements, blinks, muscle activity, and channel noise. EOG and component statistics were employed to aid artifact identification. For each participant, artifact components were removed (4.7 +/- 1.9, *group mean +/- standard deviation*) and data from remaining components back-projected into sensor space. Finally, the signal across all channels was band-pass filtered between 2-8Hz in order to extract the slow-fluctuating signals typically employed in analysis methods tracking neural responses to continuous acoustical input (Mirkovic et al., 2015; O'Sullivan et al., 2015). Following Di Liberto et al. (2015), who demonstrated that for segregating speech stimuli the envelope tracking performance with EEG was highest in the δ- (1-4Hz) and θ-bands (4-8Hz), we additionally assessed the tracking of individual frequency bands by band-pass filtering the EEG signal with FIR-filters into δ, θ, α (8-15Hz), β (15-30Hz), and γ (30-45Hz) bands.

Analysis

Behavioral Analysis & Sound Envelope Estimation

Behavioral responses were classified as hits, misses, false alarms, and correct rejections per condition, and, due to possibly differing number of trials across subjects, reported as percent accuracy. Sound onset envelopes were extracted from the music stimuli and were used in combination with EEG data to train a sound-envelope model $\boldsymbol{E}$ (*i.e.*, decoder) separately for bassoon ($\boldsymbol{E_b}$) and cello ($\boldsymbol{E_c}$; Fig. 4.2). To this end, measured EEG data was re-epoched from 2-14s to exclude activity related to both initial streaming processes or motor responses as well as any possible modulations caused by the presence of triplets in the second half of the stimulus (see Stimuli). Sound envelopes were extracted by determining the absolute Hilbert transform of each instrument independently and passing the resulting signal through a low-pass filter with a cutoff of 8Hz, of which the derivative was taken and half-wave rectified; see Hausfeld et al. (2018) for a similar approach. Such processing emphasizes short-term sound intensity fluctuations, salient in both the ongoing low-frequency EEG signals as well as in music (Hertrich et al., 2011;

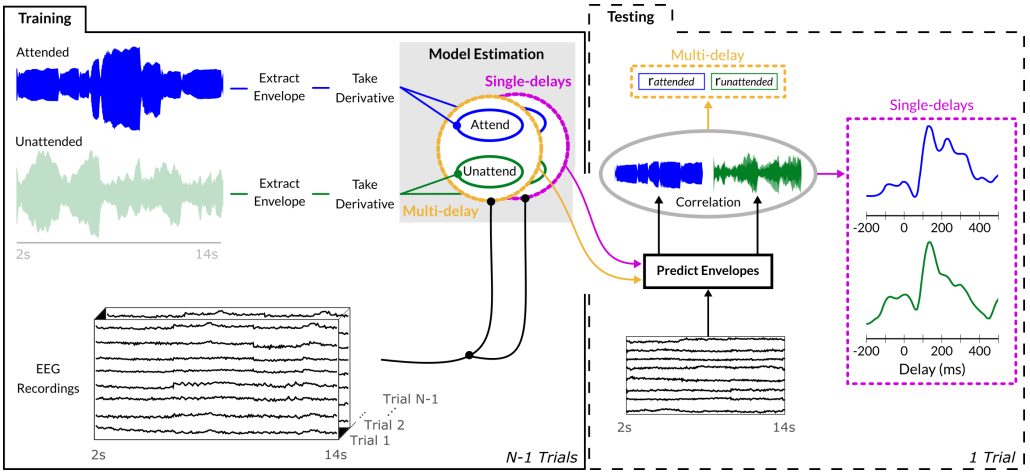


Figure 4.2: Sound Envelope Reconstruction Method. Envelopes were extracted from each instrument's waveform via an absolute Hilbert transform, its derivative was employed to estimate both single-delay and multi-delay envelope models on N-1 training trials. To assess generalization, the estimated envelope model was used to predict the sound envelope of the single unseen trial and its output correlated with the trial's actual sound envelope. Multi-delay models provided output in a single correlation value encompassing the evidence of all delays between 0-400ms (10ms step-size), while the single-delay models generated a correlation for each individual delay between -200ms to 500ms (10ms step-size).

Petersen et al., 2017; Fiedler et al., 2017); for brevity, we will refer to sound onset envelopes as sound envelopes unless specification is required.

Sound Envelope Modeling

Similar to previous EEG studies investigating envelope reconstruction (e.g., Mirkovic et al., 2015; O'Sullivan et al., 2015; Fuglsang et al., 2017), we adopted a deconvolution approach which fits, for each trial k, a multi-delay model g (*i.e.*, decoder) using the sound envelope E_k and EEG data X_k from 63 channels across 41 delays between 0-400ms (*i.e.*, 10ms step-size). The convolution kernel g_k was estimated by L2-regularized least-squares regression:

$$g_k = (X_k^T X_k + \lambda I)^{-1} X_k^T E_k \tag{4.1}$$

Regularization was performed using the identity matrix I, with the regularization parameter set to $\lambda = 10^4$ for both tasks and for all participants; this choice was based on a previous study by Hausfeld et al. (2018). The EEG data matrix X_k was constructed by concatenating the responses of all EEG channels and delays for the presented sound envelope at each individual time-point t, resulting in g_k with dimension 1201 (time points) x 2583 (channels x delays). Independent test data and sounds were employed to evaluate models on their generalization capacity to reconstruct/predict the onset envelopes from unseen bassoon, cello, or aggregate tracks ($\hat{E}_b$, $\hat{E}_c$, and $\hat{E}_a$ respectively). Model prediction and matches to sound envelopes from the test data

sets were assessed with Pearson's correlation coefficient r (Ding & Simon, 2011; Mirkovic et al., 2015; O'Sullivan et al., 2015). Generalization performance was tested within a leave-one-trial-out scheme, averaging the N-1 decoders of the training trials and applying this to the EEG data of the remaining test trial; this procedure was repeated for all trials and the correlations were averaged. The decoder g_i applied to test trial i was estimated as

$$g_i = \frac{1}{N-1} \sum_{j \neq k} g_j \tag{4.2}$$

reconstructing the unseen trial's envelope $\hat{E}_i$ by convolution

$$\hat{E}_i = g_i X_i^T \tag{4.3}$$

Envelope Model Estimation & Statistical Comparison

Within the segregation conditions, we computed models for bassoon and cello independently across all respective trials. This resulted in four different decoders: bassoon in the attention to bassoon task ($\hat{E}_b^b$), cello in the attention to bassoon task ($\hat{E}_c^b$), and *vice versa* ($\hat{E}_b^c$ and $\hat{E}_c^c$, respectively). For the aggregate condition ($\hat{E}_a$) we estimated the decoders based on the envelopes of both instruments combined.

Statistical comparisons of potential task or decoder differences for the multi-delay models were performed by virtue of non-parametric Wilcoxon signed-rank tests. In order to gain insight into those EEG delays which contribute to envelope decoding, we adopted an identical approach as above, only restricting training and testing to single delays as opposed to multiple ones. Such approach resulted in X_k representing the measurements of all channels for all time points, only at one single delay. In total 71 single-delays were tested between -200 and 500 ms; differences between models were assessed by employing a Wilcoxon signed-rank test and subsequent multiple comparison correction by a cluster-size based permutation test (Maris & Oostenveld, 2007). Specifically, we tested for each delay whether two conditions differed significantly using the Wilcoxon's sign-rank test ($p < .05$) and then summed the corresponding z-values of consecutively significant delays to obtain for each cluster its z_{sum}. These values were then compared to an empirical null-distribution obtained by permuting labels of conditions for each participant ($n_{perm} = 2^{14}$). For each permutation, clusters of significant differences were determined and the maximal z_{sum} values were extracted. This process was repeated for all permutations, each contributing a single measure to the distribution of z_{sum} values under chance given the data. Comparison of true-label values with this distribution resulted in a probability estimate corrected for multiple comparison, and those clusters which passed the $p < .05$ threshold were labeled

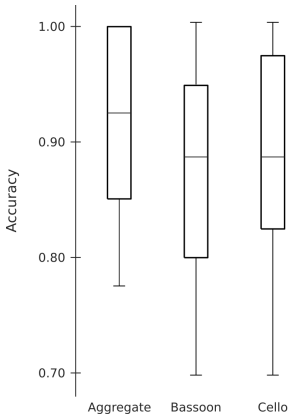


Figure 4.3: Group Behavioral Results. Accuracies across all condition trials for included subjects; box = 25th percentile - median - 75th percentile.

as significant. The EEG tracking capacity for music within frequency bands was assessed as described for multi-delay analysis, only limiting data to the δ, θ, α, β, and γ ranges.

Empirical chance level performance of the decoding models was estimated by performing the analysis as discussed, albeit with phase-scrambled versions of the stimuli ($n_{scr} = 10^4$). Such an approach keeps the frequency components of the envelopes constant. Average model performance obtained from these scrambled envelopes was compared to the non-scrambled reconstruction capacities. Note that if instrument envelopes were to be permuted instead, chance level would be overestimated due to the preservation of temporal note onsets between trials (cf. Stimulus section; Disbergen et al., 2018, Chapter 2).

In order to attempt a further disentanglement of those mechanisms at play during the integrative condition, the aggregate single-delay reconstruction time-course (r_{agg}) for each participant was fitted from a linear combination of the individual instrument reconstruction single-delay time-courses obtained during the aggregate task (r_b^a and r_c^a) using ordinary least-squares estimation:

$$r_{agg} = \beta_0 + \beta_b r_b^a + \beta_c r_c^a + \varepsilon \tag{4.4}$$

where β_0, β_b, and β_c weigh a constant and both instrument time courses, respectively.

Channel Contributions

To gain insight into which EEG channels contributed to the segregation condition's reconstruction capacity, we adopted a leave-one-channel-out approach for the single-delay models. Reconstruction of sound envelopes was achieved identically as above, only leaving one channel out for each iteration. Single-delay decoders for trial k were trained on data X_k^c (1201 [time points] x

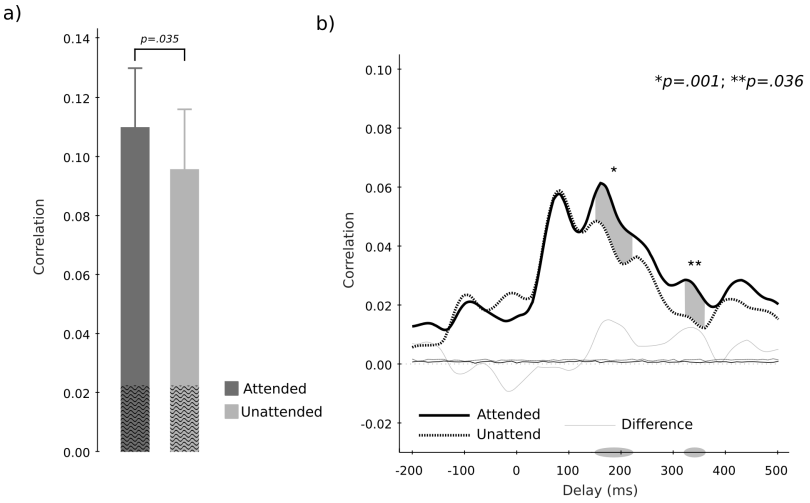


Figure 4.4: General Attended and Unattended Reconstruction Results. Multi-delay model reconstruction correlations (a) for attended (dark gray) and unattended (light gray) instruments, showing a significant difference ($p = .035$). Average empirical chance-level estimation displayed as superimposed black waves. Single-delay plots (b) displaying a significant time-resolved difference between attended (thick black solid line) or unattended (thick black dashed line) instruments during 150-220ms (*, $p = .001$) and 320-360ms (**, $p = .036$). Difference between attended and unattended reconstruction as thin gray line. Average chance-level estimations as thin solid black (attended) and thin dashed black (unattended) lines.

2542 [62 channels x 41 delays]), where c denotes the index of the left-out channel. Reconstruction correlations of the leave-one-channel-out datasets were compared with the all-channel data and their difference mapped to scalp topographies. A lower reconstruction accuracy of the left-out model results in negative values in the topographies and indicates that the respective channel possesses information relevant for the model's observed sound envelope reconstruction.

Results

Behavior

Subjects completed the experiment at high accuracy for all attention tasks: aggregate (.93 [.85 1.0], *Median [Inter Quartile Range]*), bassoon (.85 [.75 .94]), and cello (.90 [.81 .95]; Fig. 4.3). Errors resulted from a decrease in Hit rates alongside an increase in False Alarm rates, hence errors were generated equally across both of these categories, indicating there was no participant bias; comparison of false alarms from trials containing triplets in the unattended instrument versus those with no triplets present did not show differences.

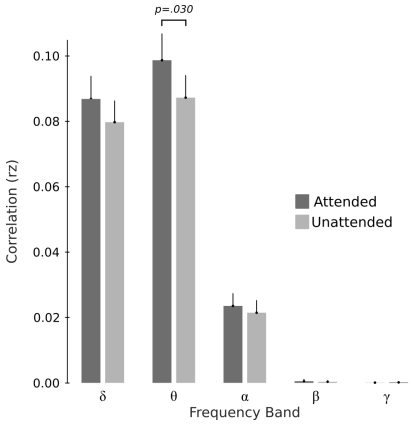


Figure 4.5: General Attended and Unattended Instrument Reconstruction Results per Frequency-band. Multi-delay model reconstruction correlations for attended and unattended instruments per EEG frequency band: delta (δ, 1-4Hz), theta (θ, 4-8Hz), alpha (α, 8-15Hz), beta (β, 15-30Hz), and gamma (γ, 30-45Hz). Showing model reconstruction capacity in the δ, θ, and α bands, with only a significant difference between attended and unattended conditions for the Theta band ($p = .030$).

Sound Envelope Tracking of Music

To examine whether there was a generalized effect on the neural representation of the attended instrument in the segregation conditions, we analyzed the data pooled across both instruments when attended vs. unattended. Correlating the envelope predictions obtained by the multi-delay models to the envelopes of test-trials, revealed that - during segregation trials - the attended ($r_z = .1069 \pm .008$; $mean \pm s.e.m.$) vs. unattended ($r_z = .096 \pm .007$) instruments displayed significantly better reconstruction of the former ($z = 2.103$, $p = .0355$, $\tilde{x}_{att-unatt} = .012$, Wilcoxon signed-rank test; Fig. 4.4a). Single-delay model correlations for the same effect indicated significantly higher reconstructions for the attended instruments at 150-220ms ($p = .001$) and at 320-360ms delay windows ($p = .036$, cluster-size based permutation test; Fig. 4.4b). Additional analysis separating the EEG data into frequency-bands indicated that, similar to speech stimuli, the tracking of attended musical instruments was best in the low frequency bands (Fig. 4.5), with better tracking for the θ- ($r_z = .099 \pm .008$) compared to the δ-band ($r_z = .087 \pm .007$; $z = 2.354$, $p = .0186$, $\tilde{x}_{\theta-\delta} = .014$). In addition, only the θ-band showed enhanced tracking of attended versus unattended instruments ($z = 2.166$, $p = .030$, uncorrected, $\tilde{x}_{att-unatt} = .011$). The tracking of instruments within higher EEG frequency bands was considerably lower (α: $r_z = .024 \pm .004$) or absent (β- and γ-band) and no attention effects were present ($p \geq .43$, uncorrected).

Additional investigations were performed into whether attended versus unattended reconstruction effects differed per instrument. Models were estimated separately for each instrument when attended or unattended, for example reconstructing bassoon during the attend to bassoon task versus the bassoon during the attend to cello task. Overall, multi-delay reconstruction

resulted in significantly higher envelope tracking for the bassoon compared to the cello instrument, both when instruments were attended ($z = 3.296$, $p < .001$, $\tilde{x}_{b-c} = .086$) or unattended ($z = 3.296$, $p < .001$, $\tilde{x}_{b-c} = .077$). For instrument tracking with multi-delay models, significantly higher tracking was found for the bassoon during the bassoon task versus the cello task ($z = 2.4797$, $p = .0132$, $\tilde{x} = .013$; Fig. 4.6a, left-hand columns), while the attention effect for the cello was not significant ($z = 1.1614$, $p = .2456$, $\tilde{x} = .010$; Fig. 4.6a, right-hand columns). Single-delay analysis showed that the bassoon was reconstructed better when attended during two delay windows, namely 170-270ms ($p = .001$) and 320-370ms ($p = .029$; Fig. 4.6b, left frame). The cello displayed a higher reconstruction correlation while attended only at delays 150-200ms ($p = .009$; Fig. 4.6b, center frame), which is comparable to the first interval for the bassoon.

The topographical contribution of EEG channels to the reconstruction of sound envelopes per instrument was obtained with a leave-one-channel-out approach, demonstrating that channels at temporal sites contributed most to the reconstruction (Fig. 4.6c). Additionally, topographies were very similar when an instrument was attended or unattended to (Fig. 4.6c).

Contrary to our hypothesis, reconstruction of aggregate time courses was not significantly better when the aggregate was the target as opposed to when only one of the instruments was attended, neither for multi-delay nor single-delay models (Fig. 4.7). Fitting of the aggregate single-delay reconstruction time-course for each participant with a linear combination of the individual instrument reconstruction time-courses obtained during the aggregate task (see Methods) suggested two distinct subgroups of subjects. A first subgroup showed similar coefficients for the bassoon (β_b) and the cello (β_c; $N = 5$; Fig. S4.2a, red stars; $\bar{\beta}_b = .608$ and $\bar{\beta}_c = .599$); a larger second group had higher coefficients for the bassoon compared to the cello ($N = 9$; Fig. S4.2a, blue squares, $\bar{\beta}_b = .752$ and $\bar{\beta}_c = .361$). Inspection of their respective single-delay aggregate condition reconstruction (Fig. S4.2b) showed significant differences between the groups during a very late delay window (400-450ms), potentially indicating distinct cognitive strategies to aggregate perception. No differences in behavioral performance were found between the two groups.

Discussion

In this work, we combined a previously validated ASA behavioral paradigm employing polyphonic music (Disbergen et al., 2018, Chapter 2) with EEG-based sound-envelope tracking methods (see Hausfeld et al., 2018, for a comparable approach) to investigate the contribution of top-down attention mechanisms to ASA. During EEG recordings, subjects were asked to detect a triplet

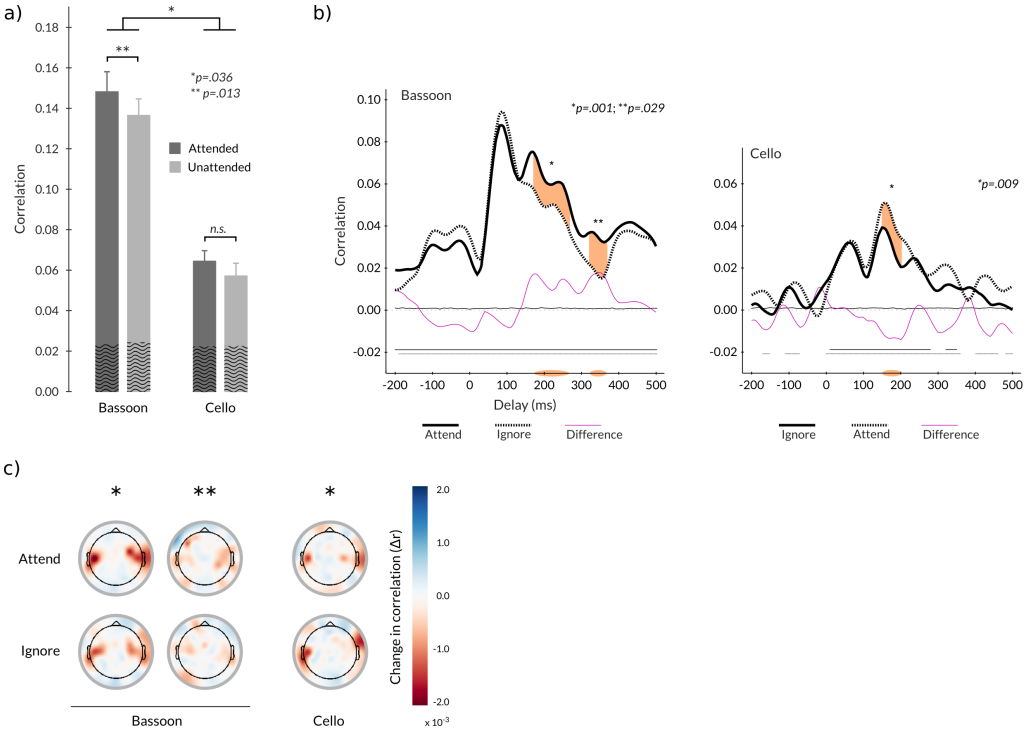


Figure 4.6: Attended and Unattended Reconstruction Results Separated per Instrument. Multi-delay model correlations (a) between reconstructed envelopes and test sounds for each condition when attended (dark gray) or unattended (light gray). Bassoon and cello reconstruction for trials trained on bassoon (left two bars) and on cello (right two bars), showing a significant difference between attended and unattended trial correlations for the bassoon instrument ($p = .013$) and not for the cello ($p = .245$). Average chance-level estimation displayed as superimposed black waves. Single-delay plots (b) showing a significant time-resolved difference for bassoon reconstruction during the 170-270ms ($p = .001$) and 320-370ms ($p = .029$) window. Cello reconstruction difference was found only during the 150-200ms window ($p = .009$). Difference between attended and unattended reconstruction as thin pink lines. Thin horizontal wavy lines within the plot indicate the average empirical chance-level estimation. Horizontal lines in the negative indicate those time-points significantly differing from chance. Topographical representation (c) of correlation change for the leave-one-electrode out analysis of both the attended and unattended conditions for each instrument during the significant delay-windows. Showing similar attended and unattended topographies for both instruments.

pattern located within or across a bassoon and cello instrument (Fig. 4.1).

Summary of the results

Results indicated that the EEG signal tracked the sound envelope of musical instruments, while across instruments the envelopes of the to-be-attended instrument was reconstructed better than those of the unattended instrument. These effects were restricted to the delay windows of 150-220ms and 320-360ms (Fig. 4.4). Separation on a per-instrument basis revealed that for multi-delay models only the envelopes of the bassoon were reconstructed better when attended to as opposed to unattended (Fig. 4.6a). In the case of time-resolved (*i.e.*, single-delay) models,

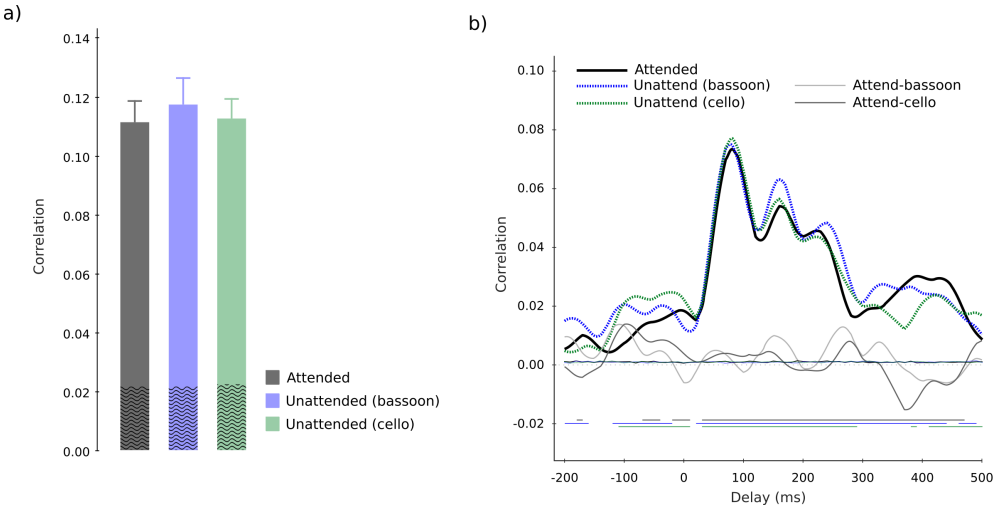


Figure 4.7: Attended and Unattended Aggregate Reconstruction Results. Multi-delay model reconstruction correlations during the aggregate condition (a) for the aggregate (gray), bassoon (blue), and cello (green), displaying no significant differences between reconstruction capacities. Average chance-level estimation displayed as superimposed black waves. Single-delay models (b) showing reconstruction capacity for the aggregate (black solid line), bassoon (blue dashed line), cello (green dashed line), and differences between aggregate and the bassoon or cello reconstruction (thin light-gray and dark-gray lines, respectively). Thin horizontal wavy lines within the plot indicate the average empirical chance-level estimation. Horizontal lines in the negative indicate those time-points significantly differing from chance.

results showed that both bassoon and cello representations were modulated by the task during a middle-latency window of 170-270ms for bassoon and 150-220ms for cello (Fig. 4.6b). In addition, we found that bassoon envelope tracking was modulated by task during an additional late-latency window of 320-370ms (Fig. 4.6b). Decoding model topographies indicated that temporal channels contributed strongest to observed effects, while topographies were similar for instruments when attended to or not (Fig. 4.6c). In contrast to our hypothesis, we did not find an attention effect for aggregate reconstruction, showing that, for both multi-delay and single-delay models, the aggregate and individual instrument envelopes were similarly reconstructed during the aggregate task (Fig. 4.7).

Stream Segregation of Instruments and Speakers

Most previous studies employing EEG-based tracking of sound-envelopes examined speech segregation in multi-speaker environments and found that physically driven mechanisms dominate effects at delays below approximately 100ms. For example, those examining temporal response functions (*e.g.*, Crosse et al., 2015), showed that initial peaks below 100ms were not modulated by attention, whereas they were by acoustical changes (Ding & Simon, 2012a,b). In addition, works examining the processing of multiple unattended sounds provided evidence that during delays below 100ms, unattended sounds remain segregated based on their acoustics,

while they get merged based on other factors only during later processing stages (Hausfeld et al., 2018; Puvvada & Simon, 2017). Consistently, the present study found that modulation by attention mainly occurred during later stages of auditory processing. In early processing windows envelope reconstruction accuracy was high but similar in the attended and unattended condition. This result is in agreement with the aforementioned speech-based ASA studies and provides a complimentary observation, suggesting similarities between speech and music regarding the early-late bisection of attentional selection. Differences between speech and music stimuli were observed mostly in the later delay windows, potentially representing top-down mechanisms from later cortical processing stages projected onto the ongoing stimulus representations.

At those time-points during which there was a significant difference between attended and unattended envelope reconstruction accuracy for individual instruments, we did not observe topographical changes when the instrument was attended to or not. The single-delay reconstruction shapes, both for general as well as instruments specific effects, were very similar between the attended and unattended condition, appearing to be enhanced when sources are attended. Taken together, these observations suggest that these effects reflect modulations of a similar cortical network, which possibly relates to the temporal-frontal network observed in our fMRI study (Chapter 3). That work showed that the listener's attended instrument could be decoded above chance at the individual subject level from the activity of frontal-temporal auditory networks, comprising large sections of the superior and medial temporal gyrus (STG, MTG), including the HG, planum polare (PP), and planum temporale (PT), sections of the inferior parietal lobe including the angular gyrus (AG), as well as varying portions of the medial and inferior frontal cortex among which the inferior frontal gyrus (IFG). Based on these observations, the attention modulations detected in the present study are potentially located in auditory cortex, originating from the medial and inferior frontal cortical regions or, alternatively, relate to more localized feedback-mechanisms within or outside of auditory areas. Described regions in the temporal-frontal network have previously been implicated in the ventral branch of the attention system, which in audition has been linked to the task-driven modulation of auditory objects (Corbetta & Shulman, 2002; Corbetta et al., 2008; Corbetta & Shulman, 2011; Cohen et al., 2009; Hill & Miller, 2010). Based on the inferior parietal lobe's role in music scene analysis, attention-related effects observed here during later delay windows could potentially contain parietal contributions which were not, or to a lesser extend, present during the earlier windows. Parietal areas have been implicated in supramodel models of top-down control and attention while performing comparable tasks (e.g., Corbetta & Shulman, 2002; Corbetta et al., 2008; Corbetta & Shulman, 2011). General task activation changes as well as integration versus segregation task differences in musicians have been observed, among others, in the intraparietal sulcus (IPS;

Janata et al., 2002; Ragert et al., 2014). The role of angular-gyrus sections has been hypothesized as an attention-modulated multisensory integrative hub (e.g., Seghier, 2012). The IPS has been hypothesized as one potential site for the origin of top-down control of auditory cortices during auditory scene analysis (e.g., Cusack, 2005; Teki et al., 2011). Within such larger-scale network, the parietal areas may also be involved in coding the temporal structures of auditory scenes (e.g., Sohoglu & Chait, 2016). In general, parietal regions appear to be part of larger-scale interactive network between auditory cortices and other non-auditory regions, associated with many complex cognitive tasks, among which musical processing.

Observation of the relatively late first occurrence of attention effects suggests that there are contributions of feedback processes to the representation and processing of music streams in a multi-instrument environment. One possible interpretation points towards a dual-stage contribution of the (early) auditory areas, a first acoustically (*i.e.*, bottom-up) driven feed-forward analysis followed by top-down feedback modulations from higher-level auditory or frontal areas. Providing there are sufficient physical differences between sounds, stimulus segregation would represent the initial feed-forward driven analysis, after which attention may interact with these ongoing bottom-up processes in these areas. Effects demonstrated here support the re-entrant activity models of stimulus representation, where active listening would modulate feedback interactions between the primary and non-primary areas, driving adaptive neuronal selection (for a review, see Gilbert & Sigman, 2007). On a network-scale, ASA probably involves a task-dependent multi-level analysis of the stimulus with a dynamic interplay between the bottom-up and, among others, attentive mechanisms (for a review, see Sussman, 2017).

Polyphonic Music Perception

Different theories of polyphonic music perception have been proposed, among which are the divided attention (Gregory, 1990) and the figure-ground model (Sloboda & Edworthy, 1981). The former suggests subjects truly divide attentional resources over the different melodic lines, while the latter poses that undivided attention is focused only on single melodic lines and polyphonic perception is achieved by rapidly alternating between melodic streams. A third dominant theory, which may potentially co-exist with the previous, suggests that listeners perform a true integration of the melodies leading to merged perception (Bigand et al., 2000). Prominent bottom-up cues which are employed in the formation of music streams are (instrument) pitch and timbre (e.g., Bregman & Pinker, 1978; Cusack & Roberts, 2000; Deutsch, 2013; Marozeau et al., 2013; McAdams, 2013a,b; Wessel, 1979). Musical notes of the same instrument are potentially first grouped based on combinations of these specific bottom-up cues, followed by interactions with top-down mechanisms.

Shapes of the single-delay reconstruction curves from the aggregate condition very much resembled that of the bassoon instrument, suggestive of a perceptual dominance for this instrument. From a music-theoretical perspective, the cello tends to be perceptually subordinate (Crawley et al., 2002; Palmer & Holleran, 1994), potentially explaining such observation. Even though we cannot provide a direct investigation into the perceptual strategies of participants, the aggregate reconstruction results hint at what subjects may potentially be employing. During aggregate reconstruction we did not find an attention effect, we did however discover two distinct subgroups of subjects with regards to their aggregate reconstruction fit regression beta-values. A first group displayed equivalent values for each instrument, while a second had a clearly stronger bassoon-weighing. One of the possible interpretations could be that these two groups differ as regards their aggregate condition perception strategies. Since the second group displayed a bassoon bias in the regression analysis, this could indicate that there was a need of more attention allocation to the cello voice, as compared to the first group. The neural representation of bassoon in group one may have been weaker compared to the subjects of group two, hence a larger beta was needed to compensate for this when fitting. When comparing the two beta-groups with regard to the shapes of their true aggregate reconstruction, we found a significant difference between them during a very late time-window (400-450ms). The late occurrence of this effect suggests it may be a very high-level cortical modulation of ongoing sound representation, which could be linked to music-specific mechanisms operating at higher-order perceptual levels (e.g., Bey & McAdams, 2002; Bregman, 1990; McAdams & Bregman, 1979).

Limitations and Considerations

No behavioral differences were found between conditions, neither here, during fMRI (Chapter 3), nor in a larger-scale behavioral study (Disbergen et al., 2018, Chapter 2). This might, however, be related to an insensitivity and/or ceiling effect of the performance metric; please see Disbergen et al. (2018, Chapter 2) for a more elaborate discussion on this as well as other task-related considerations. Across EEG analyses, we found less reconstruction capacity for models representing envelopes of the lower music voice (*i.e.*, cello), when compared to the upper music voice (*i.e.*, bassoon). This difference may be related to a general upper-voice dominance effect in the perception of polyphonic music, caused by, for example, its higher pitch (salience) or general loudness effects (Fujioka et al., 2005; Palmer & Holleran, 1994). Acoustically, there may be a continuous loudness difference between voices due to our loudness equalization method based on the overall root mean square measure, as opposed to perceptual matching. In addition, the analysis focused on rapid sound envelope fluctuations which occur more often for the bassoon as its envelope slopes are typically steeper than those of cello due to its faster attack and decay. Even though such factors may contribute to reconstruction capacity differences between

instruments, they probably did not have a great impact on discussed observations since here task-modulations of the same instrument were investigated.

In the present study, no attention effect was found for the attention to aggregate condition. Detection of such effect might be impeded by the specific task performed during the aggregate condition. Presuming the same neuronal populations represent both instruments during segregation as well as the integration tasks, the difference between both task versions may only result in very minor changes. The integrative approach could, for example, pool attentional resources more equally across those instrument-specific neuronal populations which during segregation conditions are otherwise up- and/or down-regulated. This may result in minor changes which are potentially not detectable with EEG in combination with our analysis method. We did observe a within-instrument attention effect, showing the method is sensitive to attentional changes per se, albeit the modulation effects for attending or ignoring sources are probably larger.

Due to the subject's task not requiring continuous attention allocation to the required instrument, it may be that subjects did not attend instructed instrument(s) during the full stimulus duration. Alternatively, they could have been rapidly alternating attention between the different instruments, especially in the integrative conditions. Based on previous experiments employing this paradigm, we believe that the capacity to detect triplets both within and across voices indicates that subjects were capable of segregating and integrating the instruments. Triplet detectability under both conditions provides for evidence that they managed to segregate the instruments into their individual streams. In case segregation would not have taken place, they would not have been able to respond correctly whether triplets were present within individual instruments or not. Without segregation, instruments would only differ concerning their tone on and off-sets (*i.e.*, rhythmic cues), making it impossible to assign triplets to a single voice. In general, with this paradigm we aimed at investigating which neural mechanisms permit listeners to perceive segregated or integrated melodic voices even though the physical signal arriving in their ear consists of the same identical mixed waveform under all conditions; please see Disbergen et al. (2018, Chapter 2) for additional task reflections.

Conclusion

Employing an envelope reconstruction method for EEG data, we showed that within a music ASA paradigm, attended music instruments can be significantly better reconstructed than unattended ones. Attention effects were found during delays indicative of top-down driven modulations on ongoing stimulus representations. Effects were shown both when testing a generalized attention effect across instruments as well as for the individual instruments, even though during slightly different delays. No attention effect was found for aggregate reconstruction when

compared to individual instrument reconstruction, even though two distinct subjects subgroups emerged when fitting the aggregate single-delay reconstruction time-course from a linear combination of the instrument time-courses. Discussed results extend the attentive modulation of sound envelopes in ASA into the domain of music stimuli, providing insight that similar effects previously observed with fMRI are possibly driven by top-down modulations which effect processing in the (early) auditory areas. Future research allowing for a more detailed neuronal effect localization while preserving the temporally sensitive component could shed a more detailed light onto the neuronal processes observed both here and during fMRI.

References

Alain, C. & Bernstein, L. J. (2015). Auditory Scene Analysis: Tales from Cognitive Neurosciences. *Music Perception: An Interdisciplinary Journal*, 33(1), 70–82.

Alho, K., Rinne, T., Herron, T. J., & Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: A meta-analysis of fMRI studies. *Hearing Research*, 307(c), 29–41.

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Emerson, R. G., & Schroeder, C. E. (2011). Tuning of the Human Neocortex to the Temporal Dynamics of Attended Events. *The Journal of Neuroscience*, 31(9), 3176–3185.

Bey, C. & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics*, 64(5), 844–854.

Bey, C. & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 267–279.

Bigand, E., Foret, S., & McAdams, S. (2000). Divided attention in music. *International Journal of Psychology*, 35(6), 270–278.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The Perceptual Organization of Sound. Cambridge, Massachusetts: MIT Press.

Bregman, A. S. (2015). Progress in Understanding Auditory Scene Analysis. *Music Perception: An Interdisciplinary Journal*, *33*(1), 12–19.

Bregman, A. S. & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1), 19–31.

Brochard, R., Drake, C., Botte, M. C., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1742–1759.

Carlyon, R. P. (2003). How the brain separates sounds. *Trends in Cognitive Sciences*, 8(10), 465–471.

Carlyon, R. P. & Cusack, R. (2005). Effects of Attention on Auditory Perceptual Organization. In L. Itti, G. Rees, & J. K. Tsotos (Eds.), *Neurobiology of Attention* (pp. 317–323). Cambrige, MA: Elsevier.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience-Landmark*, 13(13), 148–169.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research*, 352, 49–69.

Coffey, E. B. J., Scala, S., & Zatorre, R. J. (2011). Montreal Music History Questionnaire: a tool for the assessment of music-related experience. In *Neurosciences and Music IV Learning and Memory* Edinburgh, UK.

Cohen, Y. E., Russ, B. E., Davis, S. J., Baker, A. E., Ackelson, A. L., & Nitecki, R. (2009). A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. *Proceedings of the National Academy of Sciences*, 106(47), 20045–20050.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, 58(3), 306–324.

Corbetta, M. & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215.

Corbetta, M. & Shulman, G. L. (2011). Spatial Neglect and Attention Networks. *Annual review of neuroscience*, 34(1), 569–599.

Crawley, E. J., Acker-Mills, B. E., Pastore, R. E., & Weil, S. (2002). Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), 367–378.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, 35(42), 14195–14204.

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience*, 17, 641–651.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4), 643–656.

Cusack, R. & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, 62(5), 1112–1120.

Delorme, A. & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21.

Deutsch, D. (2013). Grouping Mechanisms in Music. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 183–248). London, UK: Elsevier.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, 25(19), 2457–2465.

Dijkstra, K. V., Brunner, P., Gunduz, A., Coon, W., Ritaccio, A. L., Farquhar, J., & Schalk, G. (2015). Identifying the attended speaker using electrocorticographic (ECoG) signals. *Brain-Computer Interfaces*, 2(4), 161–173.

Ding, N. & Simon, J. Z. (2011). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1), 78–89.

Ding, N. & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29), 11854–11859.

Ding, N. & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1), 78–89.

Disbergen, N. R., Valente, G., Formisano, E., & Zatorre, R. J. (2018). Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in Neuroscience*, 12, 70.

Disbergen, N. R., Valente, G., Zatorre, R. J., & Formisano, E. (2019). Segregation or integration of polyphonic music modulates cortical auditory responses patterns.

Fiedler, L., Wöstmann, M., Graversen, C., Brandmeyer, A., Lunner, T., & Obleser, J. (2017). Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering*, 14(3), 036020.

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage*, 156, 435–444.

Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., & Pantev, C. (2005). Automatic Encoding of Polyphonic Melodies in Musicians and Nonmusicians. *Journal of Cognitive Neuroscience*, 17(10), 1578–1592.

Gilbert, C. D. & Sigman, M. (2007). Brain States: Top-Down Influences in Sensory Processing. *Neuron*, 54(5), 677–696.

Gregory, A. H. (1990). Listening to Polyphonic Music. *Psychology of Music*, 18(2), 163–170.

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *Neuroimage*, 181, 617–626.

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2011). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology*, 49(3), 322–334.

Hill, K. T. & Miller, L. M. (2010). Auditory Attentional Control and Selection during Cocktail Party Listening. *Cerebral Cortex*, 20(3), 583–590.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective, & Behavioral Neuroscience*, 2(2), 121–140.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience*, 30(2), 620–628.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, 8(1), e53398–9.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, 77(4), 750–761.

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190.

Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, 30(3), 259–274.

McAdams, S. (2013a). Musical timbre perception. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 35–68). London, UK: Elsevier Inc.

McAdams, S. (2013b). Timbre as a structuring force in music. In *ICA 2013 Montreal* (pp. 1–6).: ASA.

McAdams, S. & Bregman, A. S. (1979). Hearing Musical Streams. *Computer Music Journal*, 3(4), 26–43.

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, 12(4), 046007.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., & Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *The Journal of Neuroscience*, 29(49), 15564–15574.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7), 1697–1706.

Palmer, C. & Holleran, S. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception & Psychophysics*, 56(3), 301–312.

Petersen, E. B., Wöstmann, M., Obleser, J., & Lunner, T. (2017). Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *Journal of Neurophysiology*, 117(1), 18–27.

Poghosyan, V. & Ioannides, A. A. (2008). Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*, 58, 802–813.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, 35(9), 1497–1503.

Puschmann, S., Baillet, S., & Zatorre, R. J. (2018). Musicians at the Cocktail Party: Neural Substrates of Musical Training During Selective Listening in Multispeaker Situations. *Cerebral Cortex*, 1537, 224–13.

Puvvada, K. C. & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *Journal of Neuroscience*, 37(38), 9189–9196.

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and Integration of Auditory Streams when Listening to Multi-Part Music. *PLoS ONE*, 9(1), 1–9.

Riecke, L., Peters, J. C., Valente, G., Kemper, V. G., Formisano, E., & Sorger, B. (2016). Frequency-Selective Attention in Auditory Scenes Recruits Frequency Representations Throughout Human Superior Temporal Cortex. *Cerebral Cortex*, advance online access, 1–13.

Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., & Desain, P. (2011). Name that tune: Decoding music from the listening brain. *Neuroimage*, 56(2), 843–849.

Seghier, M. L. (2012). The Angular Gyrus. *The Neuroscientist*, 19(1), 43–61.

Shamma, S. A. & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3), 361–366.

Sloboda, J. & Edworthy, J. (1981). Attending To Two Melodies At Once: the of Key Relatedness. *Psychology of Music*, 9(1), 39–43.

Sohoglu, E. & Chait, M. (2016). Detecting and representing predictable structure during auditory scene analysis. *eLife*, (pp. 1–17).

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *The Journal of the acoustical society of America*, 117(3), 1285–14.

Sussman, E. S. (2017). Auditory Scene Analysis: An Attention Perspective. *Journal of Speech Language and Hearing Research*, 60(10), 2989–13.

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, 69(1), 136–152.

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain Bases for Auditory Stimulus-Driven Figure-Ground Segregation. *The Journal of Neuroscience*, 31(1), 164–171.

Theunissen, F. E. & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Publishing Group*, 15(6), 355–366.

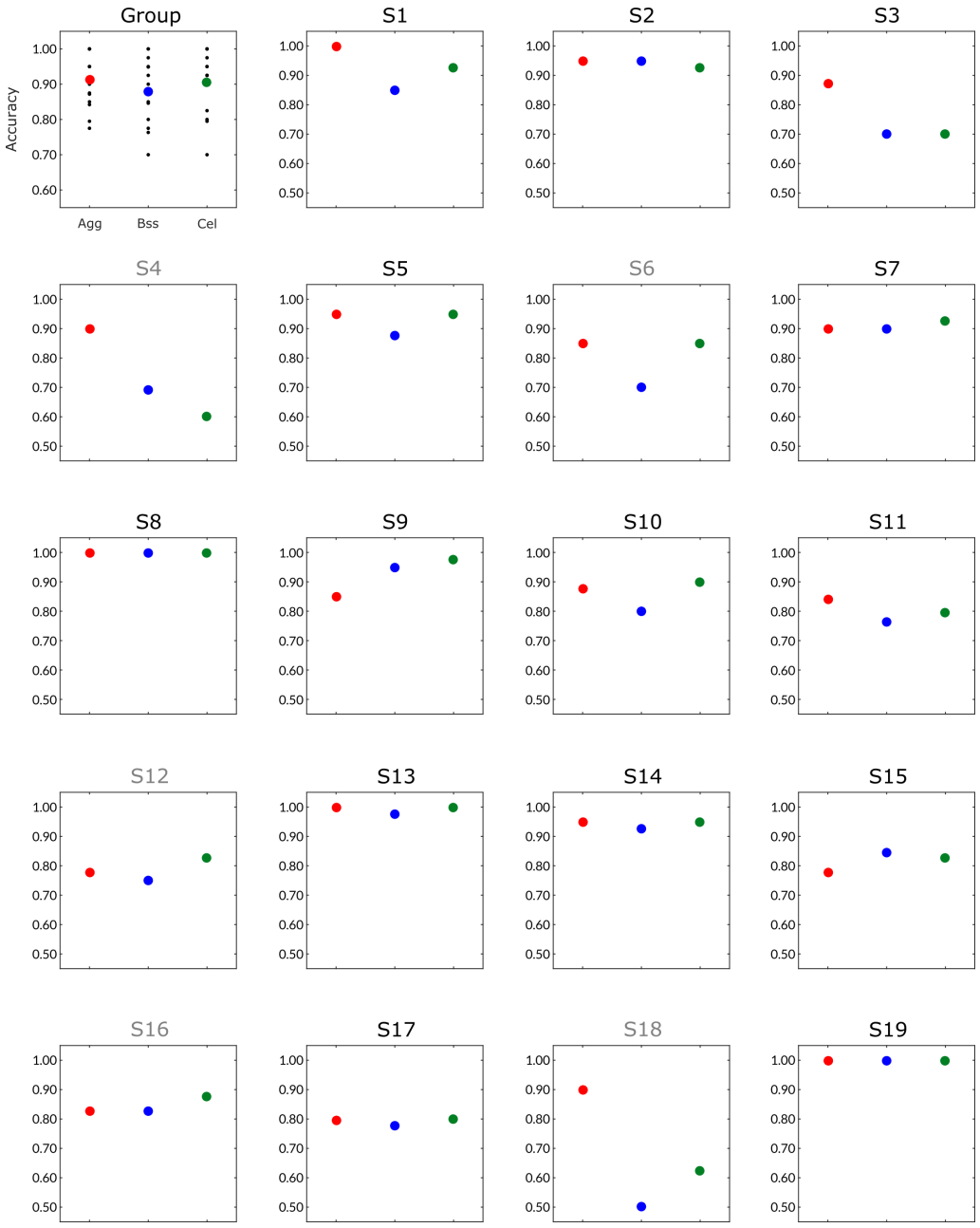
Treder, M. S., Purwins, H., Miklody, D., Sturm, I., & Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering*, 11(2), 026009–13.

Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *Neuroimage*, 77, 52–61.

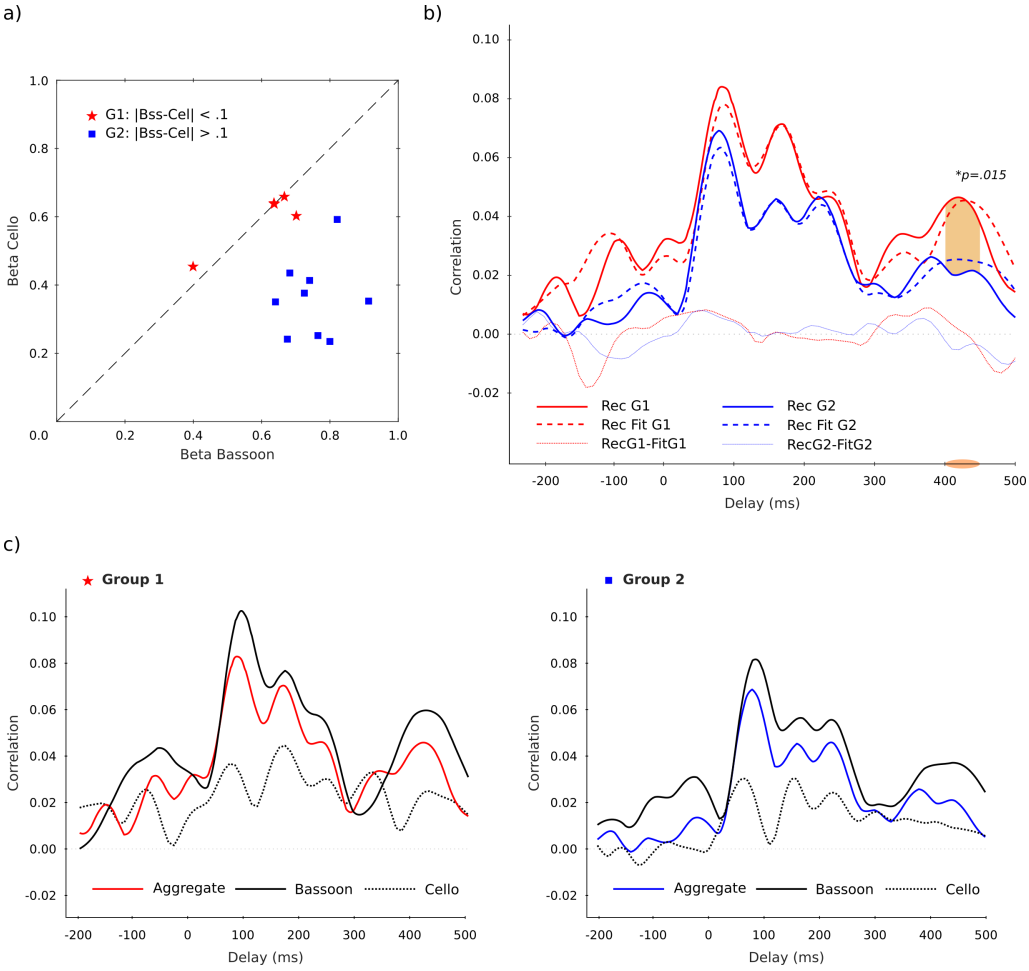
Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2), 45–52.

Woldorff, M. G. & Hillyard, S. A. (1991). Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology*, 79, 170–191.

Supplementary Figures



Supplementary Figure 4.1: Group and Individual Accuracies for all Participants. Accuracies per condition (red = aggregate, blue = bassoon, green = cello) for the group (top-left pane) and all individual subjects; excluded participant labels in gray.



Supplementary Figure 4.2: Aggregate Condition Instrument Reconstruction Results. Fitting single-delay aggregate reconstruction time courses from bassoon and cello with a linear regression (a) resulted in two distinct groups: one with a balanced beta for both (group one, red stars, N=5) and one with a larger beta for the bassoon (group two, blue squares, N=9). Single-delay model reconstruction shapes (b) of group one (solid red-line) and group two (solid blue line) differed significantly ($p = .015$) during the late delay-window 400-450ms. The aggregate reconstruction model fit per group is shown by the dashed lines, while differences between the reconstructed time courses and their model-fits are displayed by thin red (RecG1-FitG1) and blue lines (RecG2-FitG2). Reconstruction time-courses separated per group (c) for bassoon (solid black line), cello (dotted black line), and aggregate envelopes while performing the aggregate task.

5

General Discussion

Main Findings

This thesis investigated the contribution of bottom-up and top-down mechanisms to auditory scene analysis using custom composed polyphonic music, both within a behavioral and a neuroimaging setting. Two main motivations inspired the conducted studies: first, the development of a naturalistic scene analysis paradigm encompassing a behavioral metric for stimuli other than speech (chapter 2), and second, the investigation of the neural mechanisms underlying such natural complex scene analysis at high spatial (chapter 3) and temporal resolution (chapter 4).

Chapter 2 introduced an original behavioral paradigm for the investigation of ASA based on multi-instrument music. Daily ASA situations comprise an interactive interplay between bottom-up and top-down driven mechanisms. Finding a means to manipulate, in a controlled experimental setting, the relative contribution of these bottom-up and top-down mechanisms was the main goal of this study. Under the influence of attention, music can be appreciated from both a segregation and integration perspective, contrary to most other naturalistic auditory scenes. This chapter provided a behavioral validation of our paradigm in two versions: the first one investigated attentive-only (top-down) effects on the segregation and integration of music, while the second included an additional bottom-up driven modulation. During experiment 1, listeners' locus of attention was changed between integrating across both instruments (*i.e.*, integration) or attending individual instruments (*i.e.*, segregation) while detecting triplets in the music as a control for task performance. As an additional bottom-up manipulation, experiment 2 included the variation of instrument timbre distance across three discrete levels. Results of both experiments demonstrated that subjects could be trained to perform the tasks at high performance levels and there were no group differences between the attention conditions within either experiment. Experiment 2 additionally showed a main effect of instrument timbre distance, albeit failing to demonstrate this within the individual attention conditions. Intriguingly, the correlation of overall performance scores with the timbre-distance effect showed an influence of general task difficulty on the timbre-distance effect, suggesting timbre effects may be masked by task difficulty. Results confirmed that our paradigms enable the study of bottom-up and top-down driven mechanisms for auditory stream segregation and integration in both psychophysical and neuroimaging experiments.

Chapter 3 built upon the experimental paradigm introduced in chapter 2, employing it in an fMRI study at high spatial resolution which examined the cortical contributions to music ASA. Specifically, we investigated the neural correlates of top-down attentive modulations to auditory stream integration and segregation at 7 Tesla. Subjects listened to polyphonic music and

attended either the individual instruments or the aggregate while detecting the triplet patterns. Imaging data were analyzed combining independent component analysis for the unbiased definition of regions of interest with multivariate pattern classification techniques. These methods were employed to investigate the contribution of individually defined frontal-temporal auditory-responding networks to the integration and segregation of music scene elements. Results showed that the listeners' attentional state could be decoded above chance within this network, which displayed differential cortical patterns when integrating or segregating music streams. Further regional differentiation of those areas included in the temporal-frontal network showed significant above-chance classification across most of these areas, most notably already in primary auditory areas on Heschl's Gyrus. Activation patterns additionally differed when listeners attended either the bassoon or the cello, once more as early as on Heschl's Gyrus. Observations support hypotheses that indicate early auditory areas as a target of a larger-scale attentive network involved in auditory scene analysis. These early auditory areas appear to be modulated by attention in interaction with the ongoing stimulus-driven bottom-up processing mechanisms.

Chapter 4 aimed to supplement the high spatial-resolution fMRI observations of chapter 3 by adding high temporal-resolution information. This is essential for disentangling whether cortical differences observed with fMRI relate to early stimulus-driven or later top-down driven transformations of ongoing sound (feature) representations. To this end, we employed EEG in combination with the attention-only version of our paradigm (chapter 2, experiment 1), including twenty unique polyphonic pieces. EEG data were analyzed with envelope-reconstruction methods, aiming to reconstruct instrument sound-envelopes based on the measured scalp electrode data. Both single and multi-delay models showed that during the integration condition, reconstruction was not significantly modulated by attention. Vice-versa, in the instrument-segregation conditions, attended sources could be reconstructed better than unattended ones, more specifically during delays of 150-220ms and 320-360ms. Attention modulations separated per instrument only showed multi-delay model effects for the bassoon, while single-delay models indicated both bassoon and cello could be reconstructed better when attended to at middle delay windows 170-270ms and 150-200ms, respectively. Bassoon reconstruction displayed an additional late-window modulation at 320-370ms. The occurrence of attentive effects during these mid and late delay windows provided for a strong suggestion that the observed effects reflect top-down driven modulations, as opposed to purely acoustically-driven effects typically taking place at much shorter delays.

Overall, results discussed in this thesis are in favor of models suggesting contributions of auditory feedback mechanisms in the representation and processing of multiple music streams. This is in line with dual-stage models of (early) auditory cortex involvement, proposing a first feed-forward physically driven sounds analysis which is accordingly modulated by top-down

feedback arriving from later cortical processing stages, potentially modulating the feedback connectivity between primary and non-primary auditory areas (for a review, see Gilbert & Sigman, 2007). In sum, discussed results demonstrated that the integration and segregation of auditory streams resulted in different cortical representations during the analysis of naturalistic auditory scenes. They suggest that, under attentive control, separate neuronal mechanisms are driving the blending or separation of concurrent auditory streams.

Future Developments

Novel Approaches to fMRI Data Analysis

The acquisition of large individually-defined high-resolution MRI datasets at ultra-high fields opens up interesting possibilities regarding the data analysis methodology. The advantages of improved sensitivity and specificity compared to lower MRI field strengths (e.g., De Martino et al., 2018) and typical measurement error reduction (e.g., Kolossa & Kopp, 2018), have been previously discussed (see chapter 3). In addition, the availability of larger amounts of data per subject as well as multiple volumes per trial opens up new (classification) analysis possibilities. Bayesian analysis methods are an exciting venue for fMRI data analysis, especially when, like in the current situation, there are both a larger amount of data across a multitude of conditions and multiple volumes per trial.

Bayesian models in general, estimate the posterior distribution of all parameters incorporated in the model by integrating information contained in both the observations and any prior knowledge present on the problem; for an introduction to Bayesian data analysis, see Kruschke (2014). Such methods can also be employed to evaluate differences between activation patterns. Typically, the model's parameter value credibility gets more restricted when additional data is added, which results in a distribution of possible values for each model parameter that inherently describes the estimation's uncertainty. Returning to the idea behind hierarchical Bayesian models of chapter 2, it is possible to think of these activation patterns in a hierarchical fashion as well. More specifically, when a dataset comprises multiple conditions that are represented by a multitude of volumes, this naturally leads to a hierarchical model structure.

Within fMRI classification, the workhorse has for a long time been the SVM model, as employed in chapter 3, even though such a model remains limited due to the employment of two-class problems as well as single volumes (or activation estimates) per trial. The use of multiple volumes in the classification of fMRI data requires some form of data-integration, creating a merged multi-kernel which can then be used for model fitting (Fig. 5.1a), an approach which is very much

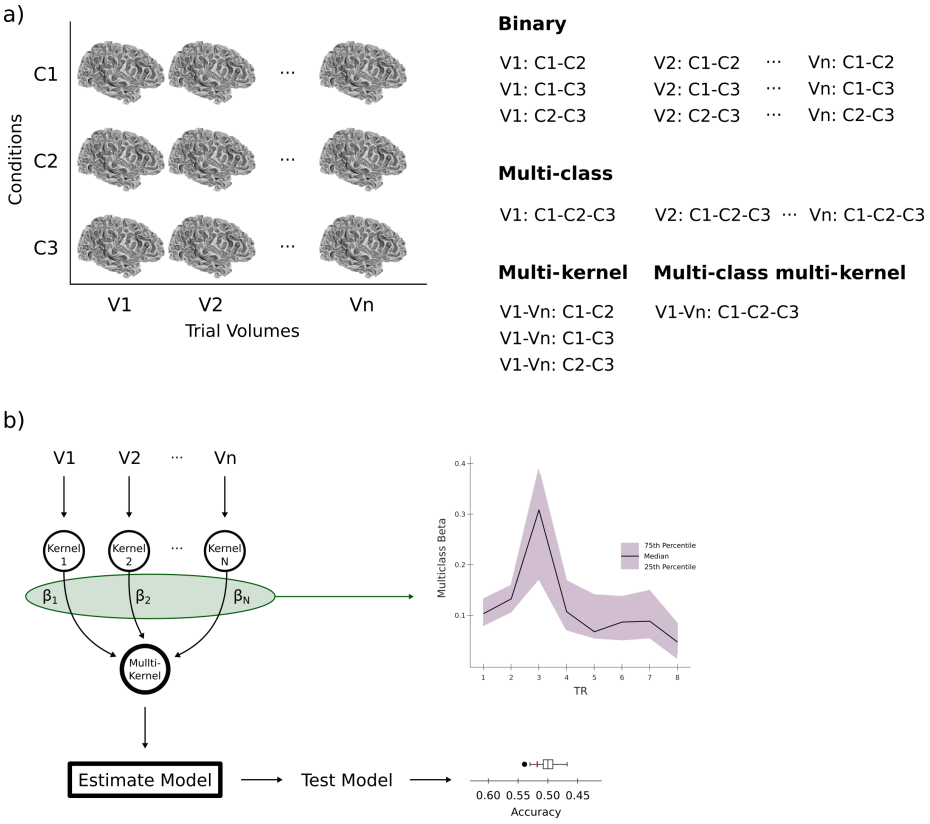


Figure 5.1: Multi-kernel and Multiclass Classification Models. When data (a) contains multiple conditions (C) and multiple volumes (V) per trial, analysis with typical binary classification models (*e.g.*, SVMs) encompasses a large number of comparisons; see right-side. Such approach limits the amount of data available for model fitting. The employment of multi-class, multi-kernel, or multi-class multi-kernel over binary models offers a great increase of available data per model. Kernels (b) can represent TRs and be combined into a multi-kernel space via linear weighing, additionally providing the possibility to extract information on TR-specific contributions to the model classification capacities.

analogous to multi-sensor fusion in engineering applications (e.g., Castanedo, 2013). Linear hierarchical kernel combinations weigh each kernel into a multi-kernel space, where each kernel would represent a TR, hence the multi-kernel weight-mapping reflects the contributions of TRs to the model's classification capacity (Fig. 5.1b). Due to the nature of multi-kernel fusion approaches, kernels can seamlessly represent different datatypes and hence potential divergence in kernel formatting due to, for example, different number of trials, will be automatically accounted for. In their rational, hierarchical Bayesian kernel combination models employed for classification (e.g., Girolami & Rogers, 2005), are very much similar to hierarchical Bayesian models such as employed in chapter 2.

When data contains a multitude of conditions in addition to multiple trial TRs, this opens up a further interesting venue for model extension and hence data-volume increase. Relevance Vector

Machines (RVMs) are of great interest in such context, as they are linear and comparable to the SVM model, even though with a Bayesian flavor and capable of seamlessly handling multiclass problems, if desired (e.g., Damoulas et al., 2008; Psorakis et al., 2010; Tipping, 2001). They are inherently sparse regression models of which classification is a special case scenario. Contrary to SVMs, they are fully probabilistic, which offers a posterior probability of class-membership for classification, which is greatly advantageous over the SVM's 'hard-boundary' decision method. It is possible to blend the hierarchical multi-kernel rational and the RVM model (e.g., Damoulas et al., 2008; Psorakis et al., 2010), leading to a potentially very powerful multi-class multi-kernel classifier. Such model is capable of informing whether classes can be disentangled in the data as well as which TRs contribute most to classification. Even though multi-class multi-kernel RVMs possess great potential for the classification of neuroimaging data, adaptations to the unique field of fMRI classification are necessary in order to successfully and meaningfully employ these models at larger scale.

Automated Cortical Network Detection

Within our current fMRI investigation, we focused on the auditory driven temporal-frontal attention network, even though a larger number of networks appears to be consistently activated across subjects. Automated detection and cross-validation of ICA-based cortical networks both within and across participants would greatly facilitate such investigations and strongly reduce the manual effort typically involved in this type of analysis (e.g., De Martino et al., 2007; Esposito et al., 2002, 2005; Himberg et al., 2004).

Voxels with consistent task-related activation should theoretically be present in the majority of the data, hence a first approach to find the most stable networks could be to split the data and search for networks present in all these analyzed splits. In the case of our experiment there is a natural split into the three experiment repetitions, hence this may form the basis of a potential within-subject network detection analysis. Identical to the chapter 3 approach, first an ICA would be performed on the active voxels GLM-Betas (Fig. 5.2). The first data-split is then used as a basis from which to start the network search, performing minimum spatial correlation selection between each base-IC and all other ICs included in the remaining data-splits. If ICs are indeed (partially) present in other data splits, this results in a (first) reduction of the possible number of matching ICs. Following initial selection, it could be checked in cross-validation whether these same spatial networks are present when the other splits form the basis from which the correlations are computed. Such an approach would result in maps representing potentially stable within-subject networks that can then be manually inspected. In order to detect likely between-subject matches of these networks within non-standardized anatomical space, the component time-courses could be correlated across subjects. Such time-course matching

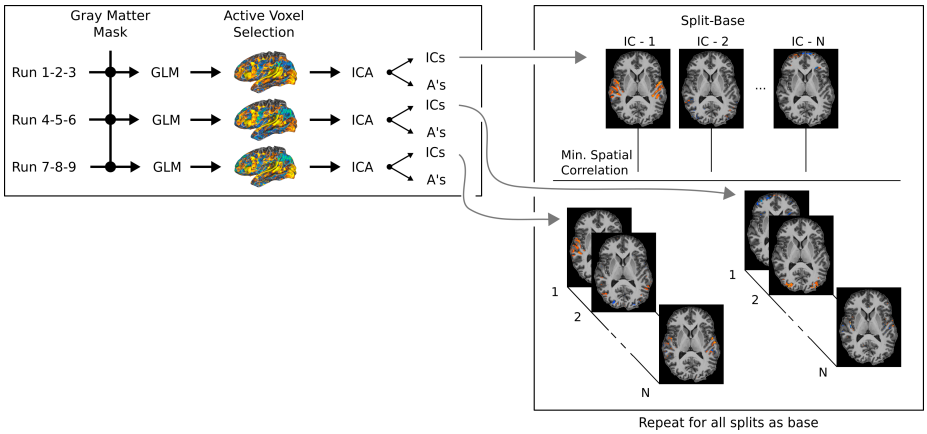


Figure 5.2: Automated Cortical Network Detection. As exemplified on the approach of chapter 3, data could be split into three experiment repetitions and ICAs estimated on their active voxel's GLM-Betas. Each of these splits ICs would accordingly be used to compute a minimal spatial correlation to all other spatial ICs from the remaining splits. Repeating this process with all the possible data-splits as a basis, allows assessment whether these spatial networks exist in all these situations.

may be performed on, for example, the raw component time-courses of voxels, or, alternatively, after computing meaningful averages, such as for experimental conditions or stimuli. This approach typically results in a large reduction of the number of stable IC-based spatial networks to be considered within subjects and could hence form the basis for further network-based analysis based on within-subject analysis methods such as discussed in chapter 3.

Further Application to the Neuroscience of Music

Here we have discussed investigations into both the behavioral and neural processes underlying ASA of long multi-instrument music stimuli. From a behavioral perspective, investigations manipulating both task and stimulus parameters provided a further in-depth understanding of both music scene analysis and its link to general ASA processes. Throughout our experiments we reached very high behavioral performance across subjects on all task-elements, as a consequence of which our behavioral metric may not be sensitive enough to detect potential subject differences in streaming performance. As such differences may be of potential interest for future (behavioral) research with our paradigm, individual changes to the stimulus parameters could be implemented to potentially introduce such sensitivity, for example by individually matching instrument pitch or timbre distance to a given performance threshold. Engagement of top-down and bottom-up processes could be further modulated by changing the pitch distance to minimum levels needed for segregation and investigating the effect instrument timbre differences have on the segregation or integration percept of listeners at this threshold. Reversely, a further reduction of instrument timbre distance could be performed by employment of instru-

ments other than bassoon and cello, for example synthesizing voices on a single instrument capable of spanning a large number of pitches or very similar within-category instruments such as a viola and cello.

Extending beyond non-musician subjects, the paradigm could be adjusted to investigate plasticity and training effects in highly trained musicians. Such application would require a modification of the triplet detection task, as in its current version it is too simplistic for assessment in this highly skilled sub-group. A potential modification could include overlapping the instrument voice pitches while synthesized on the same instrument in combination with mistuned and/or incomplete triplets within the triplet patterns. Within this highly specialized group, research could investigate potential perceptual differences between players of the various instrument sub-groups and those trained in orchestral or soloist performance.

Our neuroimaging studies provided a first insight into both the auditory-attentive brain network involved in the ASA of music and the influence locus of attention has on the cortical representations of instruments. Within the discussed MRI research, we focused on the auditory driven temporal-frontal attention network, even though a larger number of networks appears to be consistently activated across subjects. Our within-subject analysis setup employed in the fMRI experiment in conjunction with proposed automated cortical network detection could be used to further such investigations. In addition to the study of attentive modulations it is possible to disentangle effects based on instrument timbre differences included in the design. Such research would be more challenging due to the restricted number of trials available per condition, since investigations need to be performed within each attention condition to prevent biasing of timbre difference results. The number of available examples could be increased by acquiring new data focusing only on the segregation conditions while keeping three timbre-difference levels. Due to the observation of attentional modulation in early auditory areas it would be of interest to try and further disentangle these regional effects by acquiring the same experiment with MRI at a very high spatial and temporal resolution, keeping a very limited field of view including only the auditory temporal regions. Additionally, the employment of magnetoencephalography would allow to further investigate timing effects with a greatly enhanced spatial resolvability compares to EEG. Aside from design and acquisition improvements, the employment of novel analysis approaches for functional MRI data, as discussed above, could greatly enhance model sensitivity for complex data classification purposes such as encountered with our paradigms.

References

Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013(6), 704504–19.

Damoulas, T., Ying, Y., Girolami, M. A., & Campbell, C. (2008). Inferring Sparse Kernel Combinations and Relevance Vectors: An Application to Subcellular Localization of Proteins. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 577–582).: IEEE.

De Martino, F., Gentile, F., Esposito, F., Balsi, M., Di Salle, F., Goebel, R., & Formisano, E. (2007). Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. *Neuroimage*, 34(1), 177–194.

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludağ, K., De Weerd, P., Ugurbil, K., Goebel, R., & Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. *Neuroimage*, 168, 366–382.

Esposito, F., Formisano, E., Seifritz, E., Goebel, R., Morrone, R., Tedeschi, G., & Di Salle, F. (2002). Spatial independent component analysis of functional MRI time-series: To what extent do results depend on the algorithm used? *Human Brain Mapping*, 16(3), 146–157.

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., & Di Salle, F. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage*, 25(1), 193–205.

Gilbert, C. D. & Sigman, M. (2007). Brain States: Top-Down Influences in Sensory Processing. *Neuron*, 54(5), 677–696.

Girolami, M. & Rogers, S. (2005). Hierarchic Bayesian models for kernel learning.

Himberg, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, 22(3), 1214–1222.

Kolossa, A. & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *Neuroimage*, 172, 775–785.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis*. A Tutorial with R, JAGS, and Stan. London, UK: Academic Press, 2 edition.

Psorakis, I., Damoulas, T., & Girolami, M. A. (2010). Multiclass relevance vector machines - sparsity and accuracy. *IEEE Trans. Neural Networks*, 21(10), 1588–1598.

Tipping, M. E. (2001). Sparse Bayesian Learning and the relevance vector machine. *Journal of Machine Learning Research*, (pp. 1–34).

Knowledge Valorisation

During everyday listening situations we are exposed to a multitude of simultaneous sounds. Our brain processes these sounds to allow us, among other actions, to have conversations with one another or enjoy listening to music. This task becomes especially impressive when observing the signal our auditory system has to analyze: a complex pattern of air pressure changes containing information from the sound mixture. The auditory system transforms these complex mixtures into segregated sound representations to allow for the extraction of behaviorally relevant information. This thesis provided insights as to how the human brain achieves such tasks, the potential social and economical implications of which are discussed here.

Within Domain Applications

This thesis investigated the neural basis of sound perception and organization, aiming to explain the auditory system's capacity to select and respond to relevant acoustic stimuli mixed with other competing sounds. Mechanisms for music listening as have been discussed here, are of further relevance for the study of how musical training may influence general auditory cognition. Disentangling the processes underlying the Auditory Scene Analysis (ASA) of music stimuli may be fundamental to understand the neural processes supporting music processing in general (Nelken, 2008). The experimental design introduced here could be combined with other tasks which are focused on the segregative or integrative mechanisms in audition to get a better understanding of the links which exist between music-specific and general auditory cognitive abilities.

Additionally, an improved understanding of the neuronal mechanisms supporting auditory perception is of interest to other cognitive neuroscience domains, such as language and vision. Language and music have been proposed to share common processing systems, while a comparisons to vision would allow for the investigation of analogies among the senses. Research into emotion could be furthered by the employment of music stimuli as well, since they tend to evoke emotional responses in most listeners. Questions related to (music) ASA are of importance not only to cognitive neuroscience, but also neuropsychology, cognitive psychology, and, especially in this case, music psychology. Methodological advancements for MRI acquisition and analysis are transferable to other neuroscientific areas of investigation and provide the wider scientific community with novel tools to investigate brain activity.

Hearing Dysfunction and Disease

To date, most of the treatments for conductive hearing loss have focused on trying to overcome perceptual losses by virtue of sound amplitude enhancement through hearing aids. The performance of these devices is notoriously bad in noisy environments, especially so regarding speech intelligibility and music perception. The simple intra ear-canal amplification of incoming sound mixtures results in a loss or distortion of acoustical information which is essential for the auditory system to perform source segregation or integration. As a consequence of this, auditory perception may be heavily distorted. The topic of this thesis is of additional relevance to the general study of aging, since aside from this type of hearing loss the capacity to track separate auditory streams diminishes with age.

Patients employing the current hearing aid technology are often unable to perform common and socially important tasks, such as having a coffee in one's favorite cafe with a friend. Due to the many competing sound sources present in such an environment, it becomes very challenging to filter out the voice of your conversation partner while wearing a hearing aid, even though they may only be one meter away from you. Results discussed here could be of relevance to develop and enhance the general design and algorithms of both classical hearing aids and cochlear implants. An improved understanding of how the brain performs sound segregation could inform novel algorithms employing similar approaches to be implemented in these artificial hearing devices. Currently, more sophisticated algorithms are offered in hearing aids which can selectively filter and enhance relevant frequencies, in this case of speech. Unfortunately, current algorithms do not allow to overcome most of the limitations because of, among others, the erroneous 'leaking' of frequencies from other competing sound sources or general background noise.

In general, resolving mentioned hearing-aid difficulties is highly complex and potentially requires a combination of amplification modulation, selective frequency enhancement, general noise reduction, and directionality selection. If it were possible for a user to optimize their hearing-aid filtering parameters to their preference, both in general and in a situation-specific manner, via an interactive application on, for example, their smartphone, this could potentially greatly enhance their lives. One of the possible implementations could be the employment of acoustic filters which are optimized for the selection of familiar voices, introducing hearing profiles for those we interact with most. Great advancements have been made in recent years with regard to active noise canceling headphones, algorithms employed in these devices to determine what is noise and how to accordingly cancel it out could be implemented in hearing aids for noise-detection and filter adjustments. The amount of ambient noise a hearing aid perpetuates could additionally be set by the user, or automatically detected, dependent on the environment

they are in. For example, when walking around the city ambient noise is of great importance, while during a conversation it serves mostly as a distraction. Sound directionality estimations may additionally be employed to allow selection as to where the majority of sound sampling takes place, for instance on the frontal midline when having a one-on-one conversation. Considering all the possible parameter combinations which could be set on such hearing devices, it becomes of topical interest to automate them to a very large extend, preventing the need for users to spend excessive amounts of time switching between profiles or settings. Such a task requires learning a large amount of tuning parameters over time, introducing the need for learning algorithms which are capable of selecting optimal combinations of settings from a very large number of candidates, potentially in a context-dependent manner. Combining a multitude of learning architectures and algorithms, for example Bayesian reinforcement learning and data-fusion techniques, could permit the necessary combination and reduction of parameter space while continuing to allow for learning based on both algorithmic and user feedback.

Even though the research discussed here has been conducted in healthy volunteers, the advancements of its knowledge may be additionally employed to disentangle more general pathophysiological deficits of auditory cognition. For example, disorders of music cognition and perception are observed in listeners with amusia who are typically incapable of pitch perception and suffer from other deficits in both music processing and memory. Listeners with musical anhedonia show (strongly) reduced pleasure responses to music, despite having normal music perception and global functioning of the reward network. Our experimental approach may assist in developing a means to improve the characterization and diagnosis of such dysfunctions.

General Technological Applications

From a more technical perspective, general algorithm development related to stream segregation tasks, both in science and engineering, can profit from discussed advances. Both artificial speech and sound recognition relies on versions of stream segregation problems. Now that technology has become an integral part of most people's lives, there has been a great increase in efforts towards the improvement of algorithms capable of recognizing, separating, and analyzing sounds. Examples include speaker identification, speech recognition, music recognition, and most notably virtual assistants. Despite the recent phenomenal accuracy increase of these algorithms, which has been mostly driven by advancements in deep learning, these systems are far from perfect, especially in situations where there are many competing sound sources. Examples where these algorithms are heavily deployed is within the virtual assistant frameworks of Siri (Apple Inc.), Google Assistant (Google LLC), and Alexa (Amazon.com, Inc). Interaction with smart devices in both our homes and the wider environment are more and more driven through virtual assistant services, potentially becoming our main source of interaction with technology

over time. These systems make use of highly advanced algorithms, often inspired by how the brain analyses its environment. Even though we have seen great improvements, the replication of human-level performance in sound segregation and analysis still appears to be far away. Results discussed in this work could be employed in the development of novel brain-based information representation systems, potentially aiding to equate or exceed human-level performance.

References

Nelken, I. (2008). Neurons and objects: the case of auditory cortex. *Frontiers in Neuroscience*, 2(1), 107.

Publications

Articles

Disbergen, N. R., Valente, G., Formisano, E., and Zatorre, R. J. (2018). Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in Neuroscience*, 12:121.

Disbergen, N. R., Valente, G., Zatorre, R. J., and Formisano, E. (to be submitted). Segregation or integration of polyphonic music modulates cortical auditory responses patterns.

Disbergen, N. R.†, Hausfeld, L.†, Valente, G., Zatorre, R. J., and Formisano, E. (to be submitted). Modulating cortical instrument-representations during auditory stream segregation and integration with polyphonic music. † equal contribution.

Disbergen, N. R., Formisano, E., and Valente, G. (in preparation). A Bayesian multi-kernel approach to the analysis of high-resolution fMRI response patterns.

Disbergen, N. R., Valente, G., Zatorre, R. J., and Formisano, E. (in preparation). Contributions of non-auditory cortical networks to auditory stream segregation and integration with polyphonic music.

Conference Contributions

Disbergen, N. R., Valente, G., Ahrens, M., Formisano, E., and Zatorre, R. J. (2014). Streaming Music in the Brain: Development of an objective auditory stream segregation task with polyphonic music. *Fifth International Conference on Auditory Cortex (ICAC)*, September 13-17, Magdeburg, Germany.

Disbergen, N. R., Valente, G., Zatorre, R. J., and Formisano, E. (2017) Streaming Music in the Brain: Investigating the top-down modulation of auditory stream segregation and integration with polyphonic music. *Sixth International Conference on Auditory Cortex (ICAC)*, September 10-15, Banff, Alberta, Canada.

Curriculum Vitae

Niels Disbergen was born in Nijmegen, The Netherlands, on 10 May 1986. He completed his scientific high school education at the "Stedelijke Scholengemeenschap Nijmegen" in Nijmegen, The Netherlands, in 2006. During 2008 he enrolled at the Faculty of Psychology and Neuroscience of Maastricht University for a Bachelor degree in Psychology, specializing in Biological Psychology/Neuroscience. After completing his bachelor degree in 2011, he graduated in 2013 at the same faculty from the Research Master in Cognitive and Clinical Neuroscience (*cum laude*), in the Cognitive Neuroscience specialization.

During the second year of his Research Master, he spent nine months at the Montreal Neurological Institute of McGill University with Prof. Dr. Robert Zatorre to perform a research internship and write his master thesis titled "Streaming Music in the Brain: The engagement and attentive modulation of auditory stream segregation with polyphonic music". Following the successful application for a NWO Research Talent PhD-grant, in 2013 he started his PhD at the department of Cognitive Neuroscience under supervision of Prof. Dr. Formisano, Prof. Dr. Robert Zatorre, and Dr. Giancarlo Valente. During the PhD trajectory, he spent an additional six months at Prof. Zatorre's laboratory in Montreal on an Erasmus Mundus exchange grant from the European Union. After completing his PhD, he joined the design and engineering department of ASML in Veldhoven as a design engineer within the overlay reference and control group.